BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF MECHANICAL ENGINEERING
FAKULTA STROJNÍHO INŽENÝRSTVÍ

INSTITUTE OF MATHEMATICS
ÚSTAV MATEMATIKY

APPLICATION OF COUNT DATA MODELS
UŽITÍ MODELŮ DISKRÉTNÍCH DAT

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR — **Bc. Barbora Reichmanová**
AUTOR PRÁCE

SUPERVISOR — **doc. Mgr. Zuzana Hübnerová, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2018

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF MECHANICAL ENGINEERING

Specification Master's Thesis

Department:	Institute of Mathematics
Student:	**Bc. Barbora Reichmanová**
Study programme:	Applied Sciences in Engineering
Study branch:	Mathematical Engineering
Leader:	**doc. Mgr. Zuzana Hübnerová, Ph.D.**
Academic year:	2017/18

Pursuant to Act no. 111/1998 concerning universities and the BUT study and examination rules, you have been assigned the following topic by the institute director Master's Thesis:

Application of count data models

Concise characteristic of the task:

Count data models are applied in real data analysis. By the setings of the analysed experiment, the response variable has a binomial distribution with a random parameter n.

Goals Master's Thesis:

Introduction of terms and statistical tools suitable for considered problem.
Study of approaches to modeling the discrete distribution data of the studied type.
Application of statistical tools in real data analysis.

Recommended bibliography:

ANDĚL, J. Základy matematické statistiky. Praha: Matfyzpress. 2008.

DOBSON, A. J. and A. BARNETT. An Introduction to Generalized Linear Models, 3rd ed. Boca Raton: CRC Press/Taylor & Francis, 2008. ISBN 978-1-58488-950-2.

Deadline for submission Master's Thesis is given by the Schedule of the Academic year 2017/18

In Brno,

L. S.

prof. RNDr. Josef Šlapal, CSc. Director of the Institute	doc. Ing. Jaroslav Katolický, Ph.D. FME dean

Summary
When analysing the data on emergence of plants in a row of a given length, we should consider both the probability of a seed to grow successfully and a random number of seeds sown. That is why this thesis is adressing the random sums, where number of independent identically distributed summands is random number independent of the summands. The first part of the thesis focuses on the theoretical basis, where the term random sum is introduced together with various properties, numerical and functional characteristics outlining the distribution. Afterwards the method of maximum likelihood estimation is discussed, followed by generalized linear models. Moreover, the quasilikelihood method is described briefly. Throughout this part, the theory is illustrated with examples related to the initial problem. The application on real data is discussed in the last chapter.

Abstrakt
Při analýze dat růstu rostlin v řádku dané délky bychom měli uvažovat jak pravděpodobnost, že semínko zdárně vyroste, tak i náhodný počet semínek, které byly zasety. Proto se v celé práci věnujeme analýze náhodných sum, kde počet nezávisle stejně rozdělených sčítanců je na nich nezávislé náhodné číslo. První část práce věnuje pozornost teoretickému základu, definuje pojem náhodná suma a uvádí vlastnosti, jako jsou číslené míry polohy nebo funkční charakteristiky popisující dané rozdělení. Následně je diskutována metoda odhadu parametrů pomocí maximální věrohodnosti a zobecněné lineární modely. Metoda kvazi-věrohodnosti je též krátce zmíněna. Tato část je ilustrována příklady souvisejícími s výchozím problémem. Poslední kapitola se věnuje aplikaci na reálných datech a následné analýze.

Keywords
Discrete data, generalized linear models, random sums, maximum likelihood method, quasilikelihood

Klíčová slova
Diskrétní data, zobecněné lineární modely, nahodné sumy, metody maximální věrohodnosti, kvazivěrohodnost

REICHMANOVÁ, B. *Užití modelů diskrétních dat.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2018. 92 s. Vedoucí doc. Mgr. Zuzana Hübnerová, Ph.D.

Rozšířený abstrakt

Tato diplomová práce se věnuje analýze diskrétních dat, především počtu rostlin, které vyrostly ze semínka s vlivem typu orby a přítomností fungicidů. Na pokusném poli bylo vysázeno zhruba 50 semínek pomocí automatické sázečky v pěti jednometrových oddílech, proto byl počet semínek náhodný. Tento fakt vedl k analýze pomocí teorie náhodných sum.

V první kapitole zavádíme základní pojmy, které budeme v práci využívat. Definujeme charakteristickou, vytvořující a momentovou generující funkci a potřebné věty s nimi spojené. Dále uvádíme vlastnosti vybraných diskrétních rozdělení, která mohou odpovídat našemu problému. Jsou to binomické, Poissonovo, negativně binomické, diskrétní rovnoměrné a logaritmické rozdělení.

Druhá kapitola přináší diskuzi o tzv. náhodných sumách S_N, součtech nezávislých stejně rozdělených náhodných veličin $X_1, \ldots, X_N$, jejichž počet N je náhodná veličina nezávislá na X_i, tj.

$$S_N = X_1 + X_2 + \cdots + X_N.$$

Veškerá teorie je doplněna příklady vybraných diskrétních rozdělení, kde pro každou náhodnou sumu má každá náhodná veličina X_i alternativní rozdělení s pravděpodobností p. Pro takové náhodné sumy odvodíme jejich střední hodnotu a rozptyl, moment generující, vytvořující, pravděpodobnostní a charakteristickou funkci. Větší pozornost je věnována vztahu střední hodnoty a rozptylu studovaných rozdělení. Odvozená závislost je porovnána s výsledky simulací. Odvozené tvary charakteristických funkcí jsou doplněny o grafy jejich reálných a imaginárních složek s využitím `The Characteristic Function Toolbox` [3]. Ukazuje se, že je-li generující funkce veličiny N tvaru $G(\alpha(s-1))$, je typ výsledného rozdělení náhodné sumy v našem případě shodný s typem rozdělení počtu sčítanců N [2]. V našich případech se to tedy týká binomického, Poissonova a negativně binomického rozdělení. Poslední dva příklady nejsou tohoto typu, u obou je ale možné popsat výsledné rozdělení číselnými i funkčními charakteristikami.

Metoda maximální věrohodnosti je popsána v kapitole třetí. Tato metoda odhadu parametrů rozdělení je založena na principu hledání maxima tzv. věrohodnostní funkce $l(\theta; y)$, která je algebraicky totožná s pravděpodobnostní funkcí nebo hustotou, ale v tomto případě je to funkce parametru θ s fixní hodnotou y. Nejdříve zavádíme základní principy této metody pro odhadování parametrů, dále definujeme pojem rozdělení exponenciálního typu v kanonickém tvaru, jde-li její pravděpodobnostní funkci zapsat ve tvaru

$$f(y;\theta) = \exp\{yb(\theta) + c(\theta) + d(y)\},$$

kde $b(\theta), c(\theta)$ a $d(y)$ jsou známé funkce. Na konec kapitoly předkládáme příklady náhodných sum definovaných v kapitole druhé. Uvádíme zde, zda jsou exponenciálního typu, dále odvozujeme jejich věrohodnostní a logaritmickou věrohodnostní funkci a odhady parametrů. Ukazuje se, že Poissonovo, binomické (při známém n) a negativně binomické (při známém κ) rozdělení jsou exponenciálního typu. Pro model $U - Be$ je navrženo řešení numerické. Maximálně věrohodné odhady parametru v modelu $Log - Be$ lze získat transformací MLE odhadu ze článku [1].

Kapitola čtvrtá se zabývá zobecněnými lineárními modely (GLM), které umožňují analyzovat obdobu situace v lineárních regresních modelech . Dokáží popsat případy, při nichž závisle proměnné nemají normální rozdělení. Abychom mohli GLM použít, musí závisle proměnné být po dvou nezávislé a mít totožná rozdělení exponenciálního typu v kanonickém tvaru. Dalším požadavkem je lineární závislost regresorů a lineárního prediktoru $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, který je ekvivalentní hodnotě linkovací funkce $g(\mu_i)$ s proměnnou μ_i rovnou střední hodnotě i-té vysvětlované proměnné. Poté odvodíme skóre a Fisherovu míru informace pro GLM, zavádíme metody pro analýzu vhodnosti modelu a různé typy reziduí.

Následuje kapitola o kvazivěrohodnosti, které lze užít při analýze problému pro náhodnou sumu $U - Be$ a $Log - Be$. Jak je ukázáno, tato dvě rozdělení nejsou exponenciálního typu a tudíž na ně nelze aplikovat zobecněný lineární model. Proto alespoň využijeme známé vazby mezi střední hodnotou a rozptylem.

V poslední kapitole aplikujeme získané znalosti na reálných datech, převážně v programu R. Nejdříve graficky zanalyzujeme data pomocí histogramů a boxplotů, následně porovnáváme závislost rozptylu na střední hodnotě podle druhé kapitoly. Dále vytváříme zobecněné lineární modely a hledáme ten, který popisuje data nejvhodněji. Ke každému modelu uvádíme grafy popisující model a Anscombova residua.

Reference

1. GÜNEL, Erdogan a CHILKO, Daniel. Estimation of Parameter n of the Binomial Distribution. *Communications in Statistics - Simulation and Computation*. 1989, vol. 18, s. 537-551.

2. MCCULLAGH, P. a NELDER, J. A. *Generalized Linear Models*. 2. vyd. Chapman & Hall/CRC, 1989. ISBN 978-0-41-231760-6.

3. WITKOWSKY, Viktor, 2018. *The Characteristic Function Toolbox*. [online]. [Cit. 20.4.2018]. Dostupné z https://github.com/witkovsky/CharFunTool.

I hereby declare that I wrote the master thesis *Application of Count Data Models* by myself under the supervision of doc. Mgr. Zuzana Hübnerová, Ph.D. with use of the materials listed in Bibliography.

Bc. Barbora Reichmanová

I would like to express my sincerest thanks to my supervisor, doc. Mgr. Zuzana Hübnerová, Ph.D., for all the suggestions and guidance she dedicated to me, which were helpful even though we were far apart. I would like to thank prof. Correll for the data he provided us, prof. Wimmer for the help with probability generating functions, and prof. Witkowsky for sharing his programming works. I would like to thank everyone who supported and helped me during the time I worked on this thesis.

Bc. Barbora Reichmanová

Contents

Introduction

This thesis aims at an analysis of discrete data, which describes two different fields in south Australia near the city of Adelaide, Moorlands and Geranium. The seeds that were sown into the ground were 'CL Plus Grenade' that is a variety of wheat resistant to some herbicides. The soil for Moorlands had non-wetting sand and it was patchy, while on the other hand Geranium had good emergence throughout the experiment.

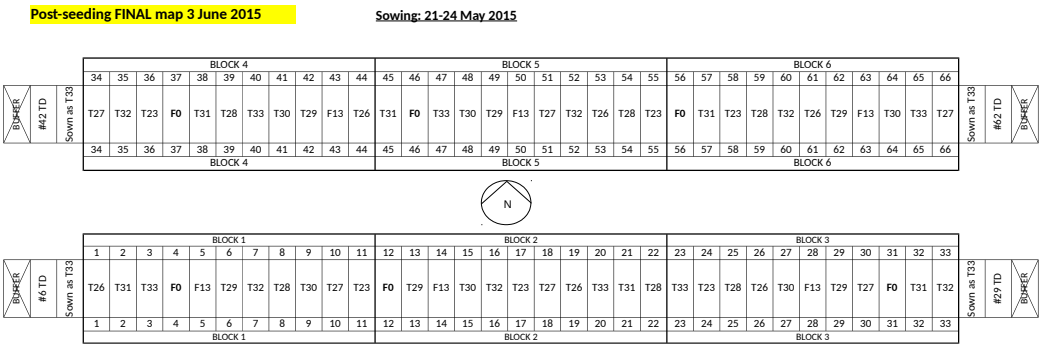


Figure 1: Moorlands

The sowing was done equally on both fields, the machine sowed around 50 seeds in five 1 meter long sections in the depth of around $4-6$ cm for Moorlands and $2-3$ cm for Geranium. It should be stressed that the number of seeds sowed is random. The variable *Plot* assigns each section a number between 1 and 66. These plots were divided into 6 blocks for each field, which corresponds to the variable *Block*. The variable *Bay* has the values 1 for northern row and 2 for southern row in the field, *Split* divides each plot into southern S and northern N part, the usage of fungicides is stored in the variable $Treated$. Last observed variable is $Treatment$ denoting the type of tillage used in the plot, in total there were 11 different types.

The final yield was about 500 kg/ha, which was 4 times smaller than initially expected. This might be due to the large effect of a severe frost.

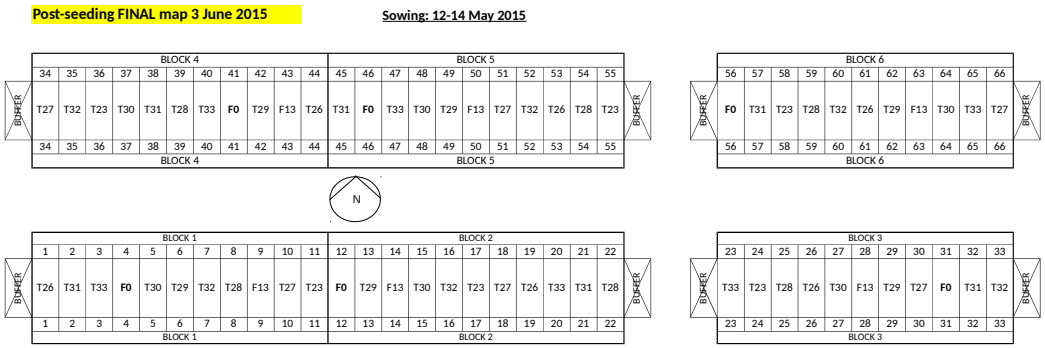


Figure 2: Geranium

The main purpose of this data analysis was to establish the effect of the type of tillage on the number of grown seeds, i.e. variable *PlantCount*. In the thesis we will first specify the theory, which might depict the problem well and then we will conduct an analysis.

In the first chapter we will lay the theoretical foundations, with which we will be working in the rest of the thesis.

The second chapter consists of a discussion about the distributions of random sums of selected discrete distributions that could characterize our problem. We derive their expected value, variance, and their relationship, moment generating, probability generating, characteristic, and probability distribution function.

The maximum likelihood method for parameter estimation is described in chapter 3. We introduce basic principles of the method and estimation, and a special case of distributions from the exponential family, which we will use in chapter 4. At the end, we present examples of the selected random sums from the previous chapter.

In the fourth chapter we define the generalized linear model and its properties. Here we use the distributions of the exponential family. We present the score and Fisher information matrix as well as the methods of goodness of fit and various residuals.

The following chapter deals with quasi-likelihood, which can be used in analysis of situations, when we cannot assess the distribution of random sum exactly as it is not from the exponential family and we are not able to apply GLM. Hence we can use the relationship between expected value and variance.

The last chapter applies gained knowledge of data analysis, mainly in the R. We compare the relationship of expected value and variance of the theoretical random sums, described in the second chapter, with the mean and variance of the data. Then we assess the generalized linear models of

various distributions, including the estimation based on quasilikelihood.

1 Terms and propositions

In this chapter we will introduce the fundamental terms and properties that will be used throughout the thesis. Only discrete case is mentioned, as we will not tackle continuous data. We were mainly using the [4], [7], [11], and [12].

Theorem 1.0.1 (Partition theorem)**.** *Let X and Y be discrete random variables. Then we have*

$$E[X] = \sum_y E[X|Y = y]\, P(Y = y).$$

Proof. The proof can be found in [12]. □

Proposition 1.0.1. *Let X and Y be two random variables. For any function g we have*

$$E[Xg(Y)|Y] = g(Y)E[X|Y]\,. \tag{1.0.1}$$

Proof. Given $Y = y$, the possible values for $Xg(Y)$ are $xg(y)$, where x varies over the range of variable X. Then

$$\begin{aligned}\mathrm{E}\,[Xg(Y)|Y = y] &= \sum_x xg(Y)P(X = x|Y = y) \\ &= g(y)\sum_x xP(X = x|Y = y) = g(y)\mathrm{E}\,[X|Y = y]\,.\end{aligned}$$

□

Theorem 1.0.2 (Law of Total Expectation)**.** *Let X be a random variable with defined expected value $E[X]$ and a random variable Y. Then*

$$E[X] = E[E[X|Y]]\,.$$

Proof. Let X and Y be discrete, then we can denote $Z = \mathrm{E}\,[X|Y]$, which is a random variable as well. Its values depend on the value y of the variable Y, i.e. $Z = \mathrm{E}\,[X|Y]$ on the set where $Y = y$. So the expected value of Z is

$$\mathrm{E}\,[Z] = \mathrm{E}\,[\mathrm{E}\,[X|Y]] = \sum_y \mathrm{E}\,[X|Y = y]\, P(Y = y) = \mathrm{E}\,[X]\,.$$

□

1.1 Functional characteristics of distributions

Definition 1.1.1. Let X be a discrete random variable, then we define the *characteristic function* of a variable X as

$$\psi_X(t) = \mathrm{E}\left[e^{itX}\right] = \mathrm{E}\left[\cos tX\right] + i\mathrm{E}\left[\sin tX\right], \qquad t \in \mathbb{R}.$$

Definition 1.1.2. The *probability generating function* of a discrete random variable X is a power series

$$G_X(s) = \mathrm{E}\left[s^X\right] = \sum_{k=0}^{\infty} p_k s^k,$$

where p_k is a probability of $X = k$, s is a complex number such that $|s| \leq 1$.

Theorem 1.1.1. *Let X and Y be independent discrete random variables. Then*

$$G_{X+Y}(s) = G_X(s) \cdot G_Y(s) \tag{1.1.1}$$

Proof. From the definition of the generating function we have that

$$G_{X+Y}(s) = \mathrm{E}\left[s^{X+Y}\right] = \mathrm{E}\left[s^X \cdot s^Y\right] = \mathrm{E}\left[s^X\right] \cdot \mathrm{E}\left[s^Y\right] = G_X(s) \cdot G_Y(s)$$

as X and Y are independent. □

Theorem 1.1.2. *If the expression exists, for a discrete random variable X we have*

$$p_0 = G_X(0),\ p_k = \frac{G_X^{(k)}(0)}{k!} \text{ for } k = 1, 2, \ldots, \tag{1.1.2}$$

where $G_X^{(k)}$ is a k-th derivative of the function G_X.

Proof. Can be found in [13]. □

Definition 1.1.3. The *moment generating function* of a discrete random variable X is

$$M_X(t) = \mathrm{E}\left[e^{tX}\right] = \sum_x e^{tx} P(X = x)$$

1.2 Distributions

Now let us introduce some discrete distributions, which we will be using further in the thesis. We have chosen the ones that could best describe our data.

1.2.1 Binomial distribution

A random variable X has the binomial distribution with the parameters n, p if its pdf is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n$$

where $p > 0$ is the probability of a success in $n \geq 1$ trials. The first two characteristics are $\mathrm{E}[X] = np$ and $\mathrm{var}[X] = np(p-1)$. We also have $\psi_X(t) = (1 - p + pe^{it})^n$, $G_X(s) = (1 - p + ps)^n$, and $M_X(t) = (1 - p + pe^t)^n$.

If $n = 1$ we call such distribution the Bernoulli distribution or Bernoulli trials.

1.2.2 Poisson distribution

A random variable X has the Poisson distribution with parameter λ if its pdf is

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1 \ldots, n$$

where $\lambda > 0$ is the expected number of occurrences. It expresses the probability of a given number of events occurring in a fixed interval, where the events occur with constant rate and independently of the last occurrence. The first two characteristics are $\mathrm{E}[X] = \mathrm{var}[X] = \lambda$. We also have $\psi_X(t) = e^{\lambda(e^{it}-1)}$, $G_X(s) = e^{\lambda(s-1)}$, and $M_X(t) = e^{\lambda(e^t-1)}$.

1.2.3 Negative binomial distribution

Let $\kappa \geq 1$ be a number of failures until the experiment is stopped and $p \in (0, 1)$ the probability of success. Then the random variable X has negative binomial distribution if its pdf is

$$P(X = x) = \binom{\kappa + x - 1}{x} p^\kappa (1-p)^x, \qquad x = 0, 1, \ldots$$

where x is the number of successes in the experiment. The first two characteristics are $\mathrm{E}[X] = \frac{p\kappa}{1-p}$ and $\mathrm{var}[X] = \frac{p\kappa}{(1-p)^2}$. We also have

$$\psi_X(t) = \left(\frac{1-p}{1-pe^{it}}\right)^\kappa,$$
$$G_X(s) = \left(\frac{1-p}{1-ps}\right)^\kappa,$$
$$M_X(t) = \left(\frac{1-p}{1-pe^t}\right)^\kappa.$$

If $\kappa = 1$ we call it the geometric distribution expressing the number of failures until the first success.

1.2.4 Discrete uniform distribution

Let $a, b \in \mathbb{Z}, a \leq b$. Then a random variable has discrete uniform distribution if its pdf is

$$P(X = x) = \frac{1}{b - a + 1} \qquad x = a, a + 1, \ldots, b.$$

The first two characteristics are $\mathrm{E}[X] = \frac{a+b}{2}$ and $\mathrm{var}[X] = \frac{(b-a+1)^2-1}{12}$. We also have

$$\begin{aligned}
\psi_X(t) &= \frac{e^{iat} - e^{i(b+1)t}}{(b - a + 1)(1 - e^{it})}, \\
G_X(s) &= \frac{s(1 - s^{b-a+1})}{(b - a + 1)(1 - s)}, \\
M_X(t) &= \left(\frac{e^{at} - e^{(b+1)t}}{1 + (b - a + 1)(1 - e^t)} \right).
\end{aligned}$$

1.2.5 Logarithmic distribution

Let $p \in (0, 1)$. Then a random variable X has the logarithmic distribution if its pdf is

$$P(X = x) = \frac{1}{-\ln(1 - p)} \frac{p^x}{x}, \qquad x \geq 1.$$

The first two characteristics are

$$\mathrm{E}[X] = \frac{-1}{\ln(1 - p)} \frac{p}{1 - p}, \qquad \mathrm{var}[X] = -p \frac{p + \ln(1 - p)}{(1 - p)^2 (\ln(1 - p))^2}.$$

We also have

$$\begin{aligned}
\psi_X &= \frac{\ln(1 - pe^{it})}{\ln(1 - p)} \\
G_X(s) &= \frac{\ln(1 - ps)}{\ln(1 - p)} \quad \text{for} \quad |s| < \frac{1}{p} \\
M_X(t) &= \frac{\ln(1 - pe^t)}{\ln(1 - p)} \quad \text{for} \quad t < -\ln(p).
\end{aligned}$$

2 Distribution of random sums

For the sake of our data, we are considering the case of a random variable S_N being a sum of independent and identically distributed discrete variables $X_1, \ldots, X_N$, where N is a discrete random variable as well. Such variable has different names across the literature random sum, compound variable or a mixture of distributions. In this chapter we will be using mainly [4], [7], [11], and [12].

Definition 2.0.1. Let $\{X_k\}$ be a sequence of mutually independent random variables with the common distribution $P(X_k = j) = f_j$ and pgf $f(s) = \sum f_j s^j$. Let us consider the sum

$$\begin{aligned} S_N &= 0 && \text{if } N = 0 \\ S_N &= X_1 + \ldots + X_n && \text{if } N = n, \end{aligned}$$

where the number N of terms is a random variable independent of the X_j with distribution $P(N = n) = g_n$ and pgf $g(s) = \sum g_n s^n$. Then the variable S_N is called the *random sum.*

The pdf of such variable is

$$h_j = P(S_N = j) = \sum_{n=0}^{\infty} P(N = n) P(X_1 + X_2 + \cdots + X_N = j).$$

Throughout this chapter we will consider the case when $X_i \sim Be(p)$ for $i = 0, 1, \ldots, n$ such that

$$\begin{aligned} \mathrm{E}[X_i] &= p \\ \mathrm{var}[X_i] &= p(p-1) \\ M_{X_i}(t) &= 1 - p + pe^t \\ G_{X_i}(s) &= 1 - p + ps \\ \phi_{X_i}(t) &= 1 - p + pe^{it}. \end{aligned}$$

2.1 Expected value and variance

Let us denote the common expected value and variance for each $i = 1, \ldots, N$

$$\mathrm{E}[X_i] = \mu \qquad\qquad \mathrm{var}[X_i] = \sigma^2.$$

Then we can determine the expected value and variance of the random sum.

Theorem 2.1.1. *The expected value of a random sum is*

$$E[S_N] = \mu E[N]. \tag{2.1.1}$$

Proof. Let us fix a nonnegative integer n, for which $N = n$. The random variable $X_1 + \ldots + X_n$ is independent of N and thereby independent of $N = n$.

$$\begin{aligned}\mathrm{E}[S_N|N = n] &= \mathrm{E}[X_1 + \ldots + X_N|N = n] = \mathrm{E}[X_1 + \ldots + X_n|N = n] \\ &= \mathrm{E}[X_1 + \ldots + X_n] = \mathrm{E}[X_1] + \ldots + \mathrm{E}[X_n] = n\mu,\end{aligned}$$

which is true for every $n \in \mathbb{N}$ and so $\mathrm{E}[S_N|N] = N\mu$. By the law 1.0.2 we get

$$\mathrm{E}[S_N] = \mathrm{E}[\mathrm{E}[S_N|N]] = \mathrm{E}[N\mu] = \mu\mathrm{E}[N].$$

□

Theorem 2.1.2. *The variance of a random sum S_N is*

$$\mathrm{var}[S_N] = \sigma^2\mathrm{E}[N] + \mu^2\mathrm{var}[N]. \tag{2.1.2}$$

Proof. To determine the variance we will use the formula

$$\mathrm{var}[S_N] = \mathrm{E}\left[S_N^2\right] - \mathrm{E}[S_N]^2.$$

Thus

$$\begin{aligned}\mathrm{E}\left[S_N^2|N = n\right] &= \mathrm{E}\left[(X_1 + \ldots + X_N)^2|N = n\right] = \mathrm{E}\left[(X_1 + \ldots + X_n)^2\right] \\ &= \mathrm{E}\left[(X_1 + \ldots + X_n)^2\right] - (\mathrm{E}[X_1 + \ldots + X_n])^2 \\ &+ (\mathrm{E}[X_1 + \ldots + X_n])^2 = \mathrm{var}[X_1 + \ldots + X_n] + n^2\mu^2 \\ &= \mathrm{var}[X_1] + \ldots + \mathrm{var}[X_n] + n^2\mu^2 = n\sigma^2 + n^2\mu^2,\end{aligned}$$

which is true for every $n \in \mathbb{N}$ and hence $\mathrm{E}[S_N^2|N] = N\sigma^2 + N^2\mu^2$. By using 1.0.2 we obtain

$$\mathrm{E}\left[S_N^2\right] = \mathrm{E}\left[\mathrm{E}\left[S_N^2|N\right]\right] = \mathrm{E}\left[N\sigma^2 + N^2\mu^2\right] = \sigma^2\mathrm{E}[N] + \mu^2\mathrm{E}\left[N^2\right].$$

Then

$$\begin{aligned}\mathrm{var}[S_N] &= \mathrm{E}\left[S_N^2\right] - (\mathrm{E}[S_N])^2 = \sigma^2\mathrm{E}[N] + \mu^2\mathrm{E}\left[N^2\right] - \mu^2(\mathrm{E}[N])^2 \\ &= \sigma^2\mathrm{E}[N] + \mu^2(\mathrm{E}\left[N^2\right] - (\mathrm{E}[N])^2) = \sigma^2\mathrm{E}[N] + \mu^2\mathrm{var}[N].\end{aligned}$$

□

Example 2.1.1 (Binomial distribution)**.** If $N \sim Bi(n, q)$ and $X_i \sim Be(p)$ then

$$\begin{aligned} \mathrm{E}\left[S_N\right] &= npq \\ \mathrm{var}\left[S_N\right] &= p(1-p) \cdot nq + p^2 \cdot nq(1-q) \\ &= \mathrm{E}\left[S_N\right] - \frac{\mathrm{E}\left[S_N\right]^2}{nq} - \frac{\mathrm{E}\left[S_N\right]^2}{n} + \frac{\mathrm{E}\left[S_N\right]^2}{nq} = \mathrm{E}\left[S_N\right] - \frac{\mathrm{E}\left[S_N\right]^2}{n}, \end{aligned}$$

which corresponds to binomial random variable with parameters n and pq.

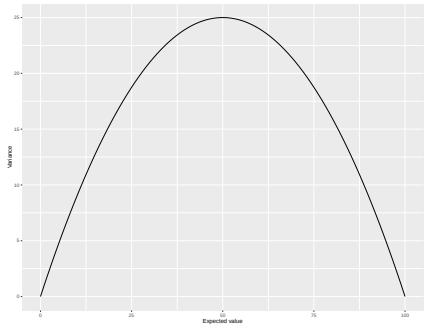


Figure 2.1: Dependence of variance on expected value of a random sum $Bi(100, q) - Be(p)$

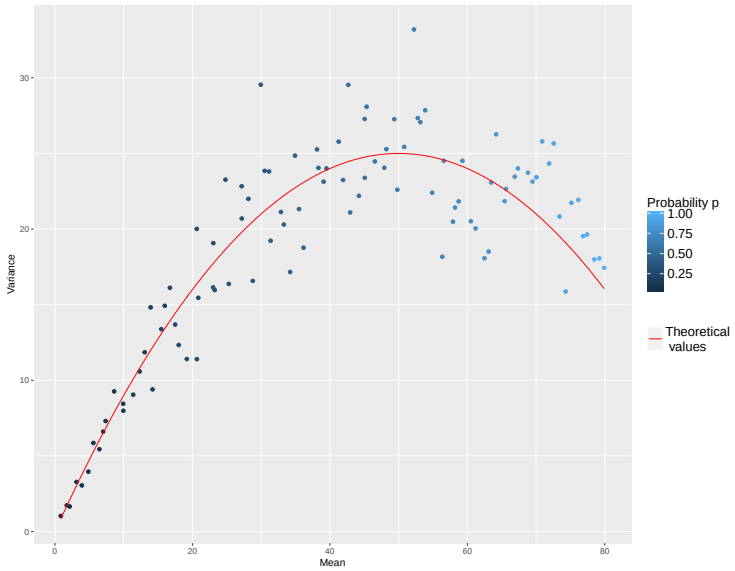


Figure 2.2: Comparison of simulation of a random sum $Bi(100, q) - Be(p)$ and theoretical values

Example 2.1.2 (Poisson distribution)**.** If $N \sim P(\lambda)$ and $X_i \sim Be(p)$ then

$$\begin{aligned} \mathrm{E}\left[S_N\right] &= p\lambda \\ \mathrm{var}\left[S_N\right] &= p(1-p)\cdot\lambda + p^2\cdot\lambda = \mathrm{E}\left[S_N\right], \end{aligned}$$

such variable corresponds to Poisson distribution with parameter $p\lambda$.

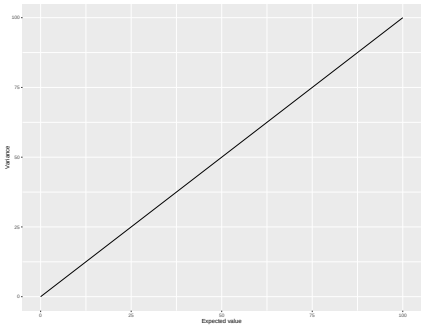


Figure 2.3: Dependence of variance on expected value of a random sum $Po(\lambda) - Be(p)$

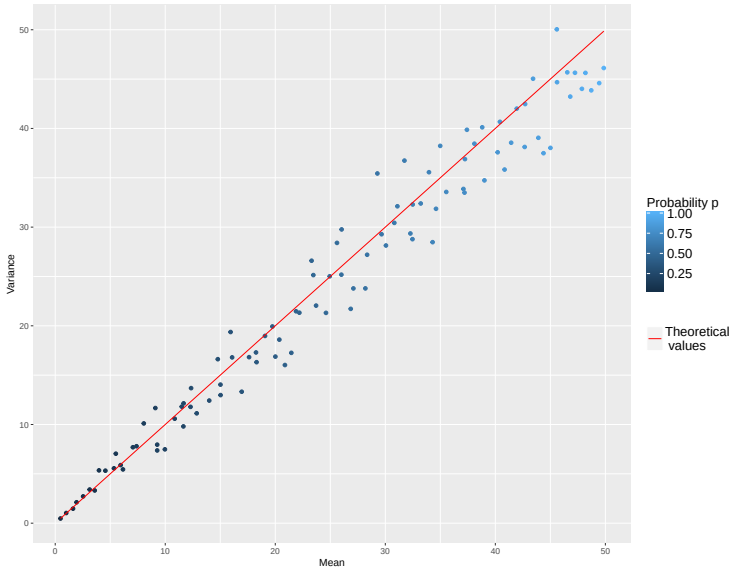


Figure 2.4: Comparison of simulation of a random sum $Po(\lambda) - Be(p)$ and theoretical values

Example 2.1.3 (Negative binomial distribution)**.** If $N \sim NBi(n, q)$ and

$X_i \sim Be(p)$, then

$$\begin{aligned}
\mathrm{E}\left[S_N\right] &= p\frac{nq}{1-q} \\
\mathrm{var}\left[S_N\right] &= p(1-p)\cdot\frac{nq}{1-q} + p^2\cdot\frac{nq}{(q-1)^2} \\
&= \mathrm{E}\left[S_N\right] - \mathrm{E}\left[S_N\right]^2\frac{1-q}{nq} + \mathrm{E}\left[S_N\right]^2\frac{1}{nq} = \frac{\mathrm{E}\left[S_N\right]^2}{n} - \mathrm{E}\left[S_N\right].
\end{aligned}$$

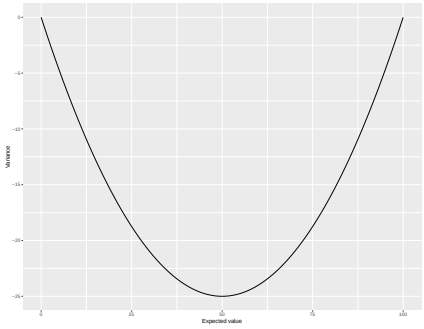


Figure 2.5: Dependence of variance on expected value of a random sum $NBi(100, q) - Be(p)$

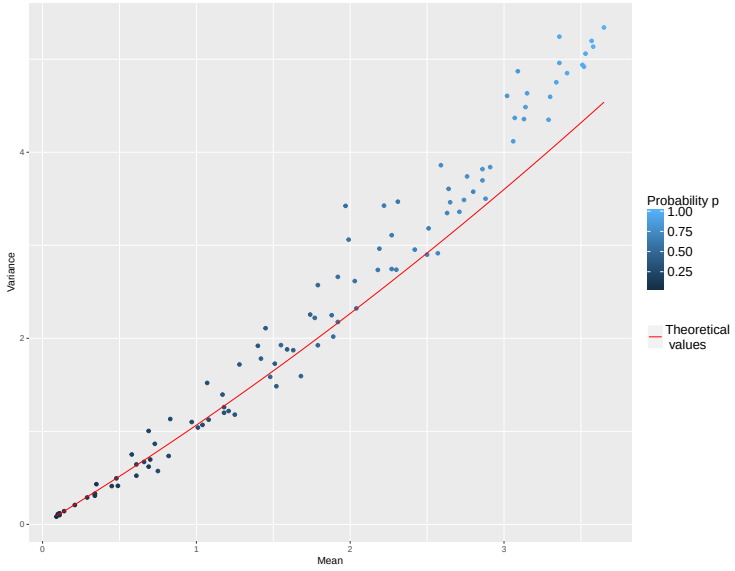


Figure 2.6: Comparison of simulation of a random sum $NBi(100, q) - Be(p)$ and theoretical values

Example 2.1.4 (Discrete uniform distribution)**.** If $N \sim U(a, b)$ and $X_i \sim Be(p)$, then

$$\begin{aligned}
\mathrm{E}\left[S_N\right] &= p \cdot \frac{a+b}{2} \\
\mathrm{var}\left[S_N\right] &= p(1-p) \cdot \frac{a+b}{2} + p^2 \cdot \frac{(b-a+1)^2 - 1}{12} \\
&= \mathrm{E}\left[S_N\right]^2 \left(-\frac{2}{a+b} + \frac{(b-a)(b-a+2)}{3(a+b)^2} \right) + \mathrm{E}\left[S_N\right].
\end{aligned}$$

For our example let us consider a special case when $a = 0$ and $b = n$, so $\mathrm{E}\left[S_N\right] = \frac{pn}{2}$ and $\mathrm{var}\left[S_N\right] = \mathrm{E}\left[S_N\right]^2 \left(-\frac{2}{n} + \frac{n+2}{3n}\right) + \mathrm{E}\left[S_N\right]$.

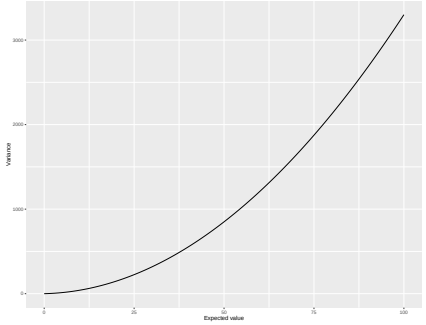


Figure 2.7: Dependence of variance on expected value of a random sum $U(0, 100) - Be(p)$

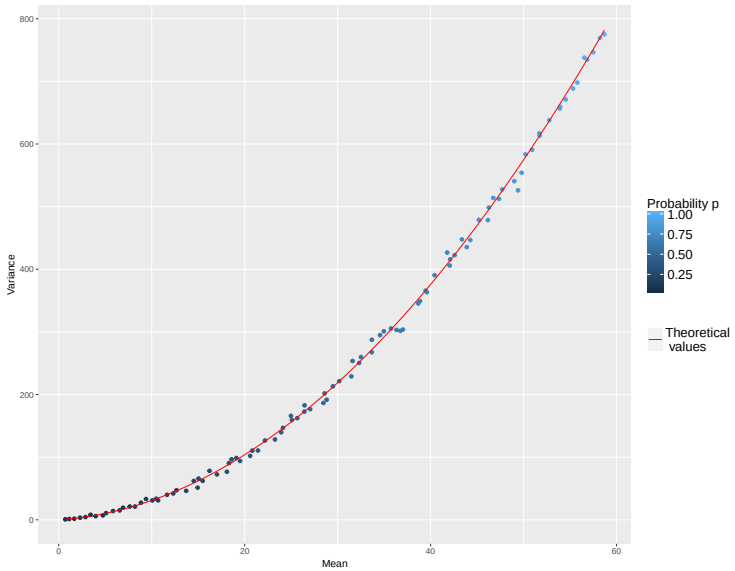


Figure 2.8: Comparison of simulation of a random sum $U(0,100) - Be(p)$ and theoretical values

Example 2.1.5 (Logaritmic distribution)**.** If $N \sim Log(q)$ and $X_i \sim Be(p)$, then

$$\begin{aligned}
\mathrm{E}[S_N] &= p \cdot \frac{-q}{\ln(1-q)(1-q)} \\
\mathrm{var}[S_N] &= p(1-p) \cdot \frac{-q}{\ln(1-q)(1-q)} + p^2 \cdot (-q) \frac{q + \ln(1-q)}{(1-q)^2 (\ln(1-q))^2} \\
&= \mathrm{E}[S_N] - \mathrm{E}[S_N]^2 (\ln(1-q) + 1).
\end{aligned}$$

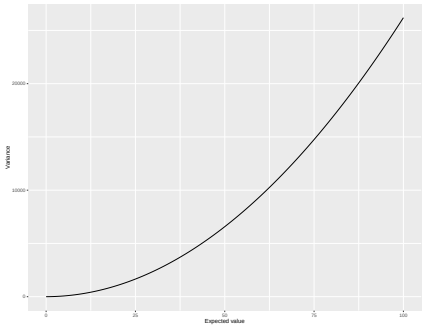


Figure 2.9: Dependence of variance on expected value of a random sum $Log(0.8) - Be(p)$

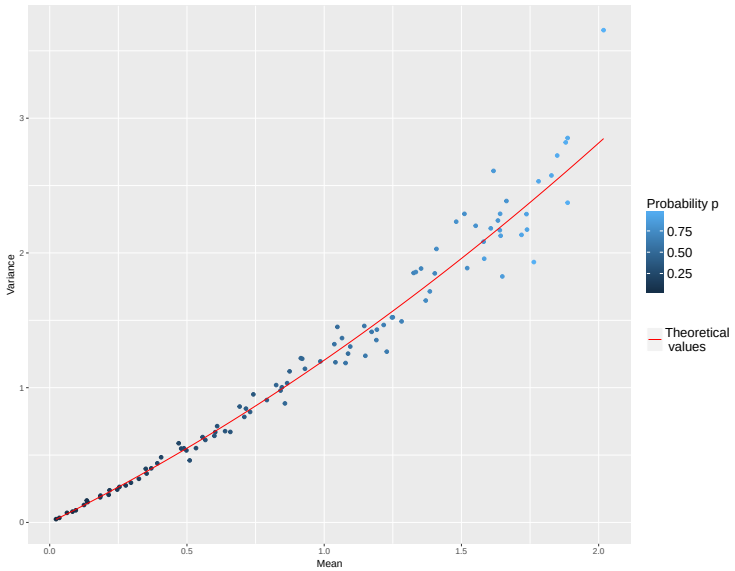


Figure 2.10: Comparison of simulation of a random sum $Log(0.8) - Be(p)$ and theoretical values

2.2 Moment generating function

Theorem 2.2.1. *Suppose we know the moment generating functions $M_N(t)$ and $M_{S_N|N}(t)$ of random variable N and S_N conditional on N. Then the unconditional moment generating function of the variable S_N is*

$$M_{S_N}(t) = M_N\left(\ln\left(M_{S_N|N}(t)\right)\right)$$

Proof.

$$\begin{aligned}
M_{S_N}(t) = \mathrm{E}\left[e^{tS_N}\right] &= \sum_y e^{ty} P(S_N = y) \\
&= \sum_y e^{ty} \sum_n P(S_N = y|N = n) \cdot P(N = n) \\
&= \sum_n \sum_y e^{ty} P(S_N = y|N = n) \cdot P(N = n) \\
&= \sum_x (M_{S_N|N}(t))^N \cdot P(N = n) \\
&= \sum_x e^{n \ln\left(M_{S_N|N}(t)\right)} \cdot P(N = n) \\
&= M_N(\ln\left(M_{S_N|N}(t)\right)).
\end{aligned}$$

□

Example 2.2.1 (Binomial distribution)**.** If $N \sim Bi(n,q)$ and $S_N|N \sim Be(p)$ we get

$$M_{S_N}(t) = (1-q+qe^{\ln((1-p+pe^t))})^n = (1-q+q(1-p+pe^t))^n = (1-pq+pqe^t)^n,$$

which is the moment generating function of the binomial distribution $S_N \sim Bi(pq,n)$ with parameter pq and n.

Example 2.2.2 (Poisson distribution)**.** If $N \sim Po(\lambda)$ and $S_N|N \sim Be(p)$ we obtain

$$M_{S_N}(t) = e^{\lambda(e^{\ln\left((1-p+pe^t)\right)}-1)} = e^{\lambda(1-p+pe^t-1)} = e^{p\lambda(e^t-1)},$$

which is the moment generating function of Poisson distribution $S_N \sim Po(p\lambda)$ with parameter $p\lambda$.

Example 2.2.3 (Negative binomial distribution)**.** If $N \sim NBi(\kappa,q)$ and $S_N|N \sim Be(p)$ we get

$$\begin{aligned} M_{S_N}(t) &= \left(\frac{1-q}{1-qe^{\ln((1-p+pe^t))}}\right)^\kappa = \left(\frac{1-q}{1-q(1-p+pe^t)}\right)^\kappa \\ &= \left(\frac{1-\frac{pq}{1-q+pq}}{1-\frac{pq}{1-q+pq}e^t}\right)^\kappa, \end{aligned}$$

which is the moment generating function of the negative binomial distribution $S_N \sim NBi(\frac{pq}{1-q+pq},\kappa)$ with the parameter κ and $\frac{pq}{1-q+pq}$.

Example 2.2.4 (Discrete uniform distribution)**.** If $N \sim U(q)$ and $S_N|N \sim Be(p)$ we obtain

$$\begin{aligned} M_{S_N}(t) &= \frac{qe^{\ln((1-p+pe^t))}}{1-(1-q)e^{\ln((1-p+pe^t))}} = \frac{q(1-p+pe^t)}{1-(1-q)(1-p+pe^t)} \\ &= \frac{q-pq+pqe^t}{p+q-pe^t-pq+pqe^t}. \end{aligned}$$

Example 2.2.5 (Logaritmic distribution)**.** If $N \sim Log(q)$ and $S_N|N \sim Be(p)$ we obtain

$$M_{S_N}(t) = \frac{\ln\left(1-qe^{\ln(1-p+pe^t)}\right)}{\ln(1-q)} = \frac{\ln(1-q(1-p+pe^t))}{\ln(1-q)}, t \leq -\ln(q).$$

2.3 Probability generating function

Let us consider a different approach concerning the generating function $G_{S_N}(s)$ of a random variable S_N as defined in 2.0.1. Then we have the following properties.

Theorem 2.3.1. *Let S_N be a random sum of the independent identically distributed variables $X_i, i = 1, \ldots, n$ with common pgf $G_X(s)$ and a random variable N being the number of summands with pgf $G_N(s)$. Then the variable S_N has the generating function*

$$G_N(G_X(s)).$$

Proof. Proof can be found in [11]. □

We will now show two types of examples. The first type has a pgf of N that can be written in the form

$$G(\alpha(s-1))$$

maybe with other parameters beside α. These underlying models are preserved under so-called Bernoulli damage process [7], i.e. in our case the random variable $X \sim Be(p)$. Such property is valid for binomial, Poisson, and negative binomial distribution of N. As will be shown, S_N then has binomial, Poisson and negative binomial distribution in sequence. Second type of examples, discrete uniform and logarithmic mixed with Bernoulli distribution, cannot be easily assessed as a known distribution.

Example 2.3.1 (Binomial distribution)**.** If $N \sim Bi(k, q)$ and $X \sim Be(p)$, then the variable S_N has the pgf

$$G_{S_N}(s) = (1 - q + q \cdot (1 - p + ps))^k = (1 - pq + pqs)^k,$$

which shows that $S_N \sim Bi(k, pq)$. We could also derive the pdf of S_N from G_{S_N} as follows. The pdf for the point $a = 0$ is

$$P(S_N = 0) = G_{S_N}(0) = (1 - pq)^k.$$

Now we compute the first and second derivative of the pgf in $s = 0$

$$G'_{S_N}(s) = k(1 - pq + pqs)^{k-1}pq$$
$$G'_{S_N}(0) = k(1 - pq)^{k-1}pq$$

$$G''_{S_N}(s) = k(k-1)(1-pq+pqs)^{k-2}p^2q^2$$
$$G''_{S_N}(0) = k(k-1)(1-pq)^{k-2}p^2q^2.$$

Then we can convey the formula for a-th derivative in 0

$$G^{(a)}_{S_N}(0) = k(k-1)\cdots(k-a+1)(1-pq)^{k-a}(pq)^a.$$

Accordingly by the theorem 1.1.2 we can derive the formula for pdf

$$P(S_N = a) = \frac{k(k-1)\cdots(k-a+1)}{a!}(1-pq)^{k-a}(pq)^a = \binom{k}{a}(pq)^a(1-pq)^{k-a}$$

which means that $S_N \sim Bi(k, pq)$ has binomial distribution with parameter k and probability pq.

Example 2.3.2 (Poisson distribution)**.** If $N \sim Po(\lambda)$ and $X \sim Be(p)$, then the pgf of a random variable S_N is

$$G_{S_N}(s) = e^{\lambda(1-p+ps-1)} = e^{\lambda p(s-1)},$$

which shows that $S_N \sim Po(p\lambda)$. From the pgf G_{S_N} we could also derive the pdf. For the point $a = 0$ the pdf is

$$P(S_N = 0) = G_{S_N}(0) = e^{-p\lambda}.$$

The first and second derivative are as follows

$$G'_{S_N}(s) = e^{\lambda p(s-1)}\lambda p$$
$$G'_{S_N}(0) = e^{-\lambda p}\lambda p$$

$$G''_{S_N}(s) = e^{\lambda p(s-1)}(\lambda p)^2$$
$$G''_{S_N}(0) = e^{-\lambda p}(\lambda p)^2.$$

Thus the formula for the a-th derivative is

$$G^{(a)}_{S_N}(0) = e^{-\lambda p}(\lambda p)^a$$

and therefore the pdf is

$$P(S_N = a) = \frac{(\lambda p)^a}{a!}e^{-\lambda p}$$

which means that $S_N \sim Po(p\lambda)$ has Poisson distribution with parameter $p\lambda$.

Example 2.3.3 (Negative binomial distribution)**.** If $N \sim NBi(\kappa, q)$ and $X \sim Be(p)$, then the pgf of the variable S_N is

$$G_{S_N}(s) = \left(\frac{1-q}{1-q\cdot(1-p+ps)}\right)^{\kappa} = \left(\frac{1-\frac{pq}{1-q+pq}}{1-\frac{pq}{1-q+pq}s}\right)^{\kappa},$$

which suggests that $S_N \sim NBi(\kappa, \frac{pq}{1-q+pq})$. Moreover, from G_{S_N} we can also derive the pdf. At the point $a = 0$ it is

$$P(S_N = 0) = G_{S_N}(0) = \left(1 - \frac{pq}{1-q+pq}\right)^{\kappa}.$$

The first and second derivative are

$$\begin{aligned} G'_{S_N}(s) &= (1-q)^{\kappa}\cdot(-\kappa)(1-q+pq-pqs)^{-\kappa-1}\cdot(-pq) \\ G'_{S_N}(0) &= (1-q)^{\kappa}\cdot(-1)^2\cdot\kappa(1-q+pq)^{-\kappa-1}\cdot pq \end{aligned}$$

$$\begin{aligned} G''_{S_N}(s) &= (1-q)^{\kappa}\cdot(-\kappa)(-\kappa-1)(1-q+pq-pqs)^{-\kappa-2}\cdot(-pq)^2 \\ G''_{S_N}(0) &= (1-q)^{\kappa}\cdot(-1)^4\cdot\kappa(\kappa+1)(1-q+pq)^{-\kappa-2}\cdot(pq)^2. \end{aligned}$$

From these we get the a-th derivative as

$$G^{(a)}_{S_N}(0) = \frac{(1-q)^{\kappa}(-1)^{2a}\cdot\kappa\cdot(\kappa+1)\cdots(\kappa+a-1)(pq)^a}{(1-q+pq)^{\kappa+a}}.$$

Finally we obtain the formula

$$\begin{aligned} P(S_N = a) &= \frac{\kappa\cdot(\kappa+1)\cdots(\kappa+a-1)}{a!}\cdot\frac{(1-q)^{\kappa}\cdot(pq)^a}{(1-q+pq)^{\kappa+a}} \\ &= \binom{\kappa+a-1}{a}\left(1-\frac{pq}{1-q+pq}\right)^{\kappa}\cdot\left(\frac{pq}{1-q+pq}\right)^{a}, \end{aligned}$$

which is the pdf of a negative binomial distribution with the parameters n and $\frac{pq}{1-q+pq}$.

Example 2.3.4 (Discrete uniform distribution)**.** If $N \sim U(0,k)$ and $X \sim Be(p)$, then the pgf of a variable S_N is

$$G_{S_N}(s) = \frac{(1-p+ps)(1-(1-p+ps)^k)}{kp(1-s)}.$$

The pdf at the point $a = 0$ is

$$P(S_N = 0) = G_{S_N}(0) = \frac{(1-p)(1-(1-p)^k)}{kp}.$$

We modify the probability generating function into a form of a sum

$$G_{S_N}(s) = \sum_{j=1}^{k} \frac{(1-p+ps)^j}{k}$$

and now we can assess the derivatives

$$G'_{S_N}(s) = \frac{1}{k}\sum_{j=1}^{k} j \cdot (1-p+ps)^{j-1} \cdot p$$

$$G'_{S_N}(0) = \frac{1}{k}\sum_{j=1}^{k} j \cdot (1-p)^{j-1} \cdot p,$$

$$G''_{S_N}(s) = \frac{1}{k}\sum_{j=1}^{k} j(j-1) \cdot (1-p+ps)^{j-2} \cdot p^2$$

$$G''_{S_N}(0) = \frac{1}{k}\sum_{j=1}^{k} j(j-1) \cdot (1-p)^{j-2} \cdot p^2.$$

From the derivatives we can obtain the formula for a-th derivative

$$G^{(a)}_{S_N}(0) = \frac{1}{k}\sum_{j=1}^{k} j(j-1)\cdots(j-a+1) \cdot (1-p)^{j-a} \cdot p^a.$$

Lastly we get the probability mass function

$$P(S_N = a) = \frac{1}{k}\sum_{j=a}^{k} \binom{j}{a} (1-p)^{j-a} p^a.$$

Example 2.3.5 (Logaritmic distribution)**.** If $N \sim Log(q)$ and $X \sim Be(p)$, then the pgf of a variable S_N is

$$G_{S_N}(s) = \frac{\ln(1-q\cdot(1-p+ps))}{\ln(1-q)} = \frac{\ln(1-q+pq-pqs)}{\ln(1-q)}, \quad |s| < \frac{1}{q}$$

The probability mass function for the point $a = 0$ is

$$P(S_N = 0) = G_{S_N}(0) = \frac{\ln(1-q+pq)}{\ln(1-q)}.$$

The first two derivatives are

$$G'_{S_N}(s) = \frac{-pq}{\ln(1-q)(1-q+pq-pqs)}$$
$$G'_{S_N}(0) = \frac{-pq}{\ln(1-q)(1-q+pq)},$$

$$G''_{S_N}(s) = \frac{(-1)\cdot(-pq)^2}{\ln(1-q)(1-q+pq-pqs)^2}$$
$$G''_{S_N}(0) = \frac{-(pq)^2}{\ln(1-q)(1-q+pq)^2}.$$

From the derivatives we can obtain the formula for a-th derivative

$$G^{(a)}_{S_N}(0) = \frac{(-1)^{2a-1}(a-1)!(pq)^a}{\ln(1-q)(1-q+pq)^a}.$$

Lastly we get the pdf

$$P(S_N = a) = \frac{(-1)^{2a-1}(a-1)!(pq)^a}{a!\ln(1-q)(1-q+pq)^a}$$
$$= \begin{cases} \frac{-1}{a\ln(1-q)}\left(\frac{pq}{1-q+pq}\right)^a & a = 1,2,\dots \\ \frac{\ln(1-q+pq)}{\ln(1-q)} & a = 0. \end{cases}$$

For more about this distribution see Example 3.3.5.

2.4 Characteristic function

In the last part of this chapter, we evaluated the characteristic function of the random sums.

Theorem 2.4.1. *The characteristic function of a random sum S_N defined in 2.0.1 is*

$$\psi_{S_N}(t) = \psi_N(-i\ln(\psi_X(t))),$$

where $\psi_N(t)$ and $\psi_X(t)$ is a characteristic function of random variable N, resp. X.

Proof. From the definition of pgf we have that $\psi_N(t) = G_N(e^{it})$ and $\psi_X(t) = G_X(e^{it})$. Then from 2.3.1

$$\psi_{S_N}(t) = G_{S_N}(e^{it}) = G_N(G_X(e^{it})) = G_N(\psi_X(t)) = G_N(e^{\ln(\psi_X(t))})$$
$$= \psi_N(i^{-1}\ln(\psi_X(t))) = \psi_N(-i\ln(\psi_X(t))).$$

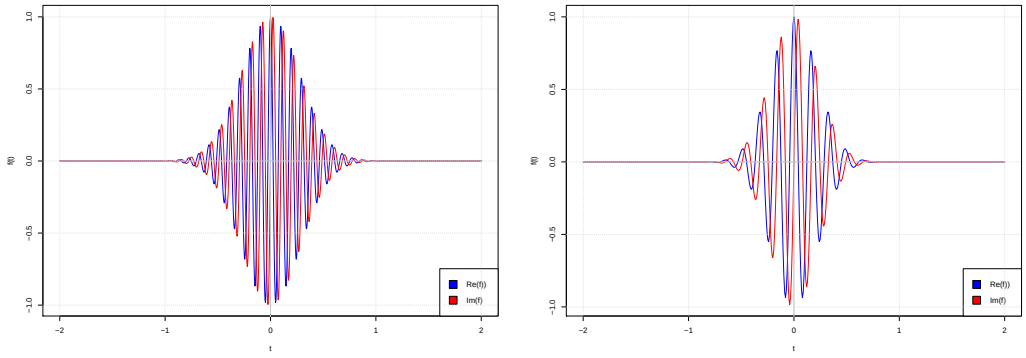
□

We also used `The Characteristic Functions Toolbox` available from [14], which let us evaluate graph of the characteristic function of a random variable, considering random sums as well.

Example 2.4.1 (Binomial distribution)**.** If $N \sim Bi(k, q)$ and $X \sim Be(p)$, then the cf of S_N is

$$\psi_{S_N} = (1 - q + qe^{i(-i\ln(1-p+pe^{it}))})^k = (1 - pq + pqe^{it})^k,$$

which is the cf of a binomial variable with the parameter k and probability pq.



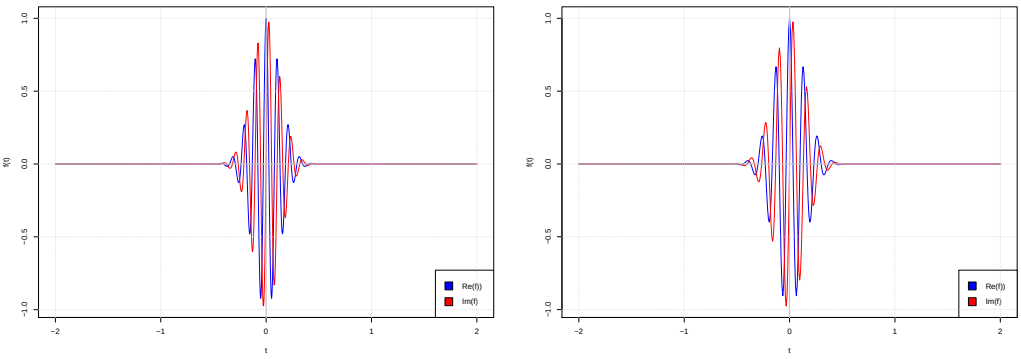
(a) Bi(80, 0.8) (b) Bi(80,0.6) - Be(0.8)

Figure 2.11: Characteristic function of Binomial variables

Example 2.4.2 (Poisson distribution)**.** If $N \sim Po(\lambda)$ and $X \sim Be(p)$, then the cf of S_N is

$$\psi_{S_N} = \exp\{\lambda(e^{i(-i\ln(1-p+pe^{it}))} - 1)\} = \exp\{\lambda p(e^{it} - 1)\};$$

which is the cf of a Poisson variable with the parameter $p\lambda$.



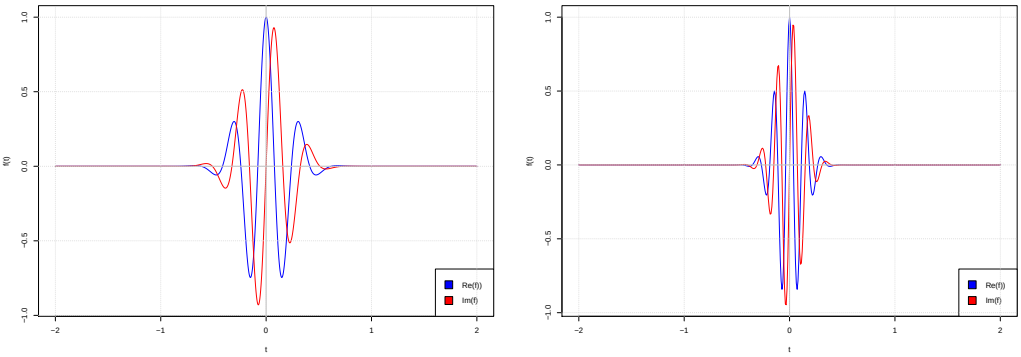
(a) Po(60) (b) Po(60) - Be(0.8)

Figure 2.12: Characteristic function of Poisson variables

Example 2.4.3 (Negative binomial distribution)**.** If $N \sim NBi(k, q)$ and $X \sim Be(p)$, then the cf of S_N is

$$\psi_{S_N} = \left(\frac{1-q}{1-qe^{i(-i\ln(1-p+pe^{it}))}} \right)^k = \left(\frac{1-\frac{pq}{1-q+pq}}{1-\frac{pq}{1-q+pq}e^{it}} \right)^k,$$

which is the negative binomial variable with parameter k and probability $\frac{pq}{1-q+pq}$.

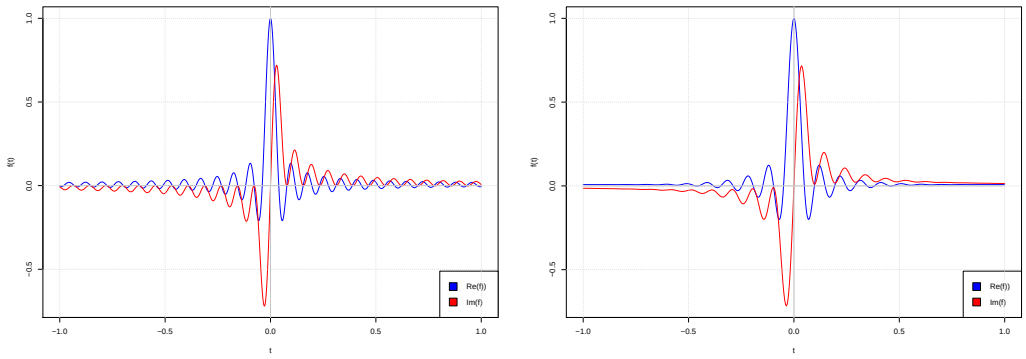


(a) NBi(80, 0.8) (b) NBi(80, 0.6) - Be(0.8)

Figure 2.13: Characteristic function of Negative binomial variables

Example 2.4.4 (Discrete uniform distribution)**.** If $N \sim U(0, k)$ and $X \sim Be(p)$, then the cf of S_N is

$$\psi_{S_N} = \frac{-e^{i(k+1)(-i\ln(1-p+pe^{it}))}}{(k+1)(1-e^{i(-i\ln(1-p+pe^{it}))})} = \frac{-(1-p+pe^{it})^{k+1}}{(k+1)p(1-e^{it})}.$$

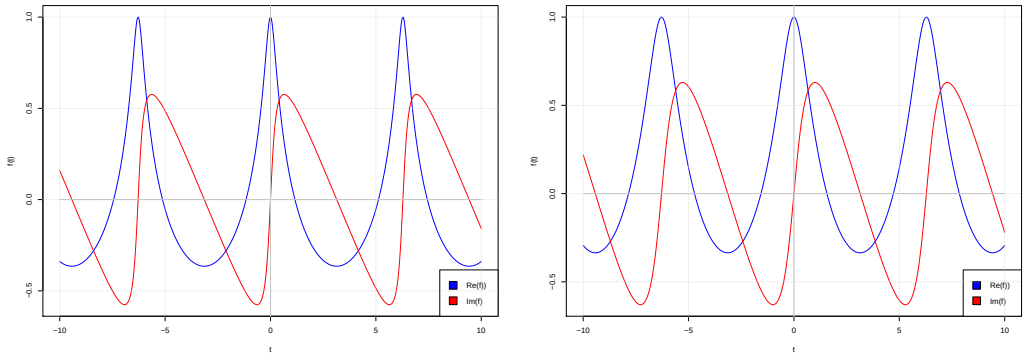


(a) U(0, 80) (b) U(0, 80) - Be(0.8)

Figure 2.14: Characteristic function of Logarithmic variables

Example 2.4.5 (Logaritmic distribution)**.** If $N \sim Log(q)$ and $X \sim Be(p)$, then the cf of S_N is

$$\psi_{S_N} = \frac{\ln\left(1 - qe^{i(-i\ln(1-p+pe^{it}))}\right)}{\ln(1-q)} = \frac{\ln(1-q+pq-pqe^{it})}{\ln(1-q)}.$$



(a) Log(0.8) (b) Log(0.6) - Be(0.8)

Figure 2.15: Characteristic function of Logarithmic variables

3 Maximum likelihood method

3.1 Basic principles

Definition 3.1.1. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be a random vector, the joint probability function $f(\mathbf{Y}; \boldsymbol{\theta})$ depends on the vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. The *likelihood function* $L(\boldsymbol{\theta}; \mathbf{Y})$ is algebraically the same as the joint probability but the emphasis shifted from the random variable $\mathbf{Y}$ with $\boldsymbol{\theta}$ fixed to the parameters $\boldsymbol{\theta}$ with $\mathbf{Y}$ fixed.

The *log-likelihood function* is the likelihood function $l(\boldsymbol{\theta}; \mathbf{Y}) = \log L(\boldsymbol{\theta}; \mathbf{Y})$.

Because the function L is defined in terms of the random variable $\mathbf{Y}$, it itself is a random variable.

Definition 3.1.2. Let Ω denote the set of all possible values of the parameter vector $\boldsymbol{\theta}$, then Ω is called the *parameter space*.

Definition 3.1.3. Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be a random vector with pdf $f(\mathbf{x}, \boldsymbol{\theta})$ with respect to some σ-finite measure μ. Suppose the following properties hold:

(a) $\boldsymbol{\theta} \in \Omega$, where Ω is non-empty open set in $\mathbb{R}^p$,

(b) the set $M = \{\boldsymbol{x} : f(\mathbf{x}, \boldsymbol{\theta}) > 0\}$ is independent of $\boldsymbol{\theta}$,

(c) for almost all $\mathbf{x} \in M$ with respect to μ and for all $i = 1, \ldots, p$ there exist partial derivatives $f_i'(\mathbf{x}, \boldsymbol{\theta}) = \frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_i}$,

(d) $\int_M f_i'(\mathbf{x}, \boldsymbol{\theta}) \, \mathrm{d}\mu(\mathbf{x}) = 0$ for all i and for all $\boldsymbol{\theta} \in \Omega$,

(e) For every pair (i, j) the integral

$$J_{ij}(\boldsymbol{\theta}) = \int_M \frac{f_i'(\mathbf{x}, \boldsymbol{\theta}) f_j'(\mathbf{x}, \boldsymbol{\theta})}{f^2(\mathbf{x}, \boldsymbol{\theta})} f(\mathbf{x}, \boldsymbol{\theta}) \, \mathrm{d}\mu(\mathbf{x})$$

is finite,

(f) The matrix $\mathfrak{J}_n(\boldsymbol{\theta}) = \|J_{ij}(\boldsymbol{\theta})\|_{i,j=1}^{p}$ is positive definite for every $\boldsymbol{\theta} \in \Omega$.

Then the system of pdf's $\{f(\mathbf{x}, \boldsymbol{\theta}), \boldsymbol{\theta}\}$ is called *regular* and $\mathfrak{J}_n(\boldsymbol{\theta})$ is called the *Fisher information matrix*.

The gradient of log-likelihood function

$$\mathbf{U}(\boldsymbol{\theta}) = \nabla l(\boldsymbol{\theta}, \mathbf{Y}) = \left(\frac{\partial l(\boldsymbol{\theta}, \mathbf{Y})}{\partial \theta_1}, \ldots, \frac{\partial l(\boldsymbol{\theta}, \mathbf{Y})}{\partial \theta_p} \right)$$

is called the *score*.

Theorem 3.1.1. *Let f be the pdf from a regular system of pdf's $\mathcal{F}$ and let the second partial derivatives of the pdf $f''_{ij}(\mathbf{Y}, \boldsymbol{\theta})$ exist. Then the score vector $\mathbf{U}$ corresponding to the pdf f has the expected value $E[\mathbf{U}] = 0$ and the variance-covariance matrix $var[\mathbf{U}] = \mathcal{J}$. Moreover,*

$$E\left[\frac{f''_{ij}(\mathbf{Y}, \boldsymbol{\theta})}{f(\mathbf{Y}, \boldsymbol{\theta})} \right] = 0, \quad i, j = 1, \ldots, n,$$

then $var[\mathbf{U}] = -E[\mathbf{U}'_{\boldsymbol{\theta}}]$.

Proof. Can be found in [6]. □

Definition 3.1.4. The *maximum likelihood estimator* of the parameter $\boldsymbol{\theta}$ is the value $\hat{\boldsymbol{\theta}}$ which maximizes the likelihood function, that is

$$L(\hat{\boldsymbol{\theta}}; \mathbf{Y}) \geq L(\boldsymbol{\theta}; \mathbf{Y}) \qquad \forall \boldsymbol{\theta} \in \Omega.$$

Equivalently, $\hat{\boldsymbol{\theta}}$ maximizes the log-likelihood function, as the logarithmic function is monotonic.

The algorithm to find an estimator $\hat{\boldsymbol{\theta}}$ is same as finding a maximum of a function, therefore, if the derivatives exist, we firstly need to solve the differential equations

$$\frac{\partial l(\boldsymbol{\theta}; \mathbf{Y})}{\partial \theta_j} = 0 \qquad \text{for} \quad j = 1, \ldots, p. \tag{3.1.1}$$

We need to check the second derivatives, namely that the matrix $\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}^2}$ evaluated at the point $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ is negative definite, i.e.

$$\left[\frac{\partial^2 l(\boldsymbol{\theta}; \mathbf{Y})}{\partial \boldsymbol{\theta}^2} \right]_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} < 0.$$

Moreover, it is necessary to check values of $\boldsymbol{\theta}$ on the border of Ω that give the local maxima of $l(\boldsymbol{\theta}; \mathbf{Y})$. The value of $\boldsymbol{\theta}$ corresponding to the largest value of the function $l(\boldsymbol{\theta}; \mathbf{Y})$ is the maximum likelihood estimator.

Definition 3.1.5. Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be a random vector with the pdf $f(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Omega \subset \mathbb{R}^p$. Let $u : \Omega \to \Omega^* \subset \mathbb{R}^k$ and $G(\boldsymbol{\theta}^*) = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \Omega, u(\boldsymbol{\theta}) = \boldsymbol{\theta}^*\}$. Let us denote

$$M(\mathbf{x}, \boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta} \in G(\boldsymbol{\theta}^*)} f(\mathbf{x}, \boldsymbol{\theta}).$$

Then the function M of parameter $\boldsymbol{\theta}^*$ is called *likelihood function induced by parametric function* u. The value $\boldsymbol{\theta}^*$ maximizing $M(\mathbf{X}, \boldsymbol{\theta}^*)$ is called *maximum likelihood estimator of the parametric function* u.

Theorem 3.1.2 (Zehna's theorem - invariance principle for maximum likelihood estimators)**.** *If $\boldsymbol{\theta}^*$ is maximum likelihood estimator of the parameter $\boldsymbol{\theta}$ then $u(\hat{\boldsymbol{\theta}})$ is maximum likelihood estimator of the parametric function $u(\boldsymbol{\theta})$.*

Proof. Can be found in [2]. □

3.2 Exponential family of distributions

Definition 3.2.1. Let Y be a random variable, whose distribution depends on a single parameter θ. The distribution belongs to the *exponential family in canonical form* if its pdf can be written as

$$f(y; \theta) = \exp\left[yb(\theta) + c(\theta) + d(y)\right], \tag{3.2.1}$$

where $b(\theta), c(\theta)$, and $d(y)$ are known functions.

We want to determine the expected value and variance of a variable with regular distribution from the exponential family. We can derive the score

$$U = Yb'(\theta) + c'(\theta).$$

From Theorem 3.1.1 we have that $\mathrm{E}[U] = 0$, i.e.

$$\mathrm{E}[Y] = -\frac{c'(\theta)}{b'(\theta)}. \tag{3.2.2}$$

Because $U' = Yb''(\theta) + c''(\theta)$, we have

$$\mathrm{var}[U] = -\mathrm{E}[U'] = -b''(\theta)\mathrm{E}[Y] - c''(\theta) = b''(\theta)\frac{c'(\theta)}{b'(\theta)} - c''(\theta).$$

On the other hand $\mathrm{var}[U] = (b'(\theta))^2 \mathrm{var}[Y]$ and therefore

$$\mathrm{var}[Y] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}. \tag{3.2.3}$$

From this we have

$$\mathfrak{J} = \mathrm{var}[U] = [b'(\theta)]^2 \mathrm{var}[Y] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{b'(\theta)}. \tag{3.2.4}$$

3.3 Examples

As the last part of this chapter we will show examples of distributions we were using in chapter 2. The first three distributions, binomial, Poisson, negative binomial, are of the exponential family, whereas the latter two, discrete uniform - Bernoulli, logarithmic - Bernoulli, are not. In both cases we will show the form of the likelihood or log-likelihood function, the functions $b(\theta), c(\theta)$, and $d(y)$ for distributions of the exponential family, and the estimation of the parameter $\hat{\theta}$, if they exist.

Example 3.3.1 (Binomial distribution)**.** Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be a random sample of size n from $Bi(k, p)$ with k known parameter, then the likelihood and log-likelihood function is

$$L(p; \mathbf{Y}) = \prod_{i=1}^{n} \exp\{y_i(\ln(p) - \ln(1-p)) + k\ln(1-p) + \ln\binom{k}{y_i}\}$$

$$l(p; \mathbf{Y}) = \ln(L(p; \mathbf{Y})) = \ln\left(\frac{p}{1-p}\right)\sum_{i=1}^{n} y_i + kn\ln(1-p) + \sum_{i=1}^{n}\ln\binom{k}{y_i}.$$

The distribution is regular and from the exponential family in the canonical form, we have that $b(p) = \ln\left(\frac{p}{1-p}\right)$, $c(p) = kn\ln(1-p)$, and $d(\mathbf{Y}) = \sum_{i=1}^{n}\ln\binom{k}{y_i}$. Therefore the score and Fisher information is

$$U = \frac{\sum_{i=1}^{n} Y_i - kpn}{p(1-p)}$$

$$\mathfrak{J} = \frac{k}{p(1-p)}.$$

Now we can determine the estimate of p from the equation

$$\frac{\partial l}{\partial p} = \frac{\sum_{i=1}^{n} Y_i - kpn}{p(1-p)} = 0$$

$$\Rightarrow \hat{p} = \frac{\sum_{i=1}^{n} Y_i}{kn}.$$

Example 3.3.2 (Poisson distribution)**.** Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be a random sample of size n from $Po(\lambda), i = 1, \ldots, n$, then the likelihood and log-likelihood function is

$$L(\lambda; \mathbf{Y}) = \prod_{i=1}^{n} \exp\{y_i\ln(\lambda) - \lambda - \ln(y_i!)\}$$

$$l(\lambda; \mathbf{Y}) = \ln(\lambda)\sum_{i=1}^{n} y_i - n\lambda - \sum_{i=1}^{n}\ln(y_i!).$$

The distribution is regular and from the exponential family in canonical form, we have that $b(\lambda) = \ln(\lambda)$, $c(\lambda) = -n\lambda$, and $d(\mathbf{Y}) = -\sum_{i=1}^{n} \ln(y_i!)$. The score and Fisher information is

$$U = \frac{\sum_{i=1}^{n} Y_i}{\lambda} - n$$
$$\mathfrak{J} = \lambda.$$

Now we can determine the estimate of λ from the equation

$$\frac{\partial l}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} Y_i - n = 0$$
$$\Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

Example 3.3.3 (Negative binomial distribution)**.** Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be a random sample of size n from $NBi(k, p), i = 1, \ldots, n$ with k known parameter, then the likelihood and log-likelihood function is

$$L(p; \mathbf{Y}) = \prod_{i=1}^{n} \binom{y_i + k - 1}{y_i} p^{y_i} (1-p)^k$$
$$l(p; \mathbf{Y}) = \ln(L(p; \mathbf{Y})) = \ln(p) \sum_{i=1}^{n} y_i + kn \ln(1-p) + \sum_{i=1}^{n} \ln \binom{y_i + k - 1}{y_i}.$$

The distribution is regular and from the exponential family in the canonical form, we have that $b(p) = \ln(p)$, $c(p) = kn \ln(1-p)$, and $d(\mathbf{Y}) = \sum_{i=1}^{n} \ln \binom{y_i+k-1}{y_i}$. Therefore the score and Fisher information is

$$U = \frac{(1-p)\sum_{i=1}^{n} Y_i - pkn}{p(1-p)}$$
$$\mathfrak{J} = \frac{k}{(1-p)^2}.$$

Now we can determine the estimate of p from the equation

$$\frac{\partial l}{\partial p} = \frac{(1-p)\sum_{i=1}^{n} Y_i - pkn}{p(1-p)} = 0$$
$$\Rightarrow \hat{p} = \frac{\sum_{i=1}^{n} Y_i}{nk + \sum_{i=1}^{n} Y_i}.$$

Example 3.3.4 (Discrete uniform - Binomial distribution)**.** Regarding the form of the pdf derived in the Chapter 3 we are not able to find the estimates easily. It can be solved numerically as a optimization problem of finding a maximum of a function, for instance in R.

Example 3.3.5 (Logarithmic - Binomial distribution)**.** This random sum is described by [9] and [7] and is called the extension of a logarithmic series or logarithmic-with-zeros distribution. If we have $N \sim Log(q)$ and $X \sim Be(p)$ then we can reparametrize the pdf of S_N as

$$P(S_N = y) = \begin{cases} \omega & \text{for } y = 0 \\ -\frac{(1-\omega)\theta^y}{y \ln(1-\theta)} & \text{for } y = 1, 2, \ldots, \end{cases}$$

where the parameters ω and θ are transformation of p and q as

$$\omega = \frac{\ln(1 - q + pq)}{\ln(1 - q)}$$
$$\theta = \frac{pq}{1 - q + pq}.$$

The ML estimators of ω and θ are given by

$$\begin{aligned} \hat{\omega} &= \frac{f_0}{n} \\ \sum_{j \geq 1} \frac{j f_j}{n} &= \frac{-\hat{\theta}}{(1 - \hat{\theta}) \ln\left(1 - \hat{\theta}\right)}, \end{aligned} \tag{3.3.1}$$

where f_j is the observed frequency of an observation equal to j. Using 3.1.1 the estimation of parameter q can be obtained from 3.3.1.

4 Generalized linear models

Consider linear models of the form

$$E(Y_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}_i; \qquad Y_i \sim N(\mu_i, \sigma^2), \tag{4.0.1}$$

where Y_i are independent identically distributed random variables, the vector $\mathbf{x}_i^T$ represents the i-th row of the design matrix $\mathbf{X}$. In this section, we want to focus on cases generalizing the linear models, where response variables have distributions other than Normal (they can be either continuous or discrete) and the relationship between response and explanatory variables does not need to be linear. In this chapter we will be using [1], [3], and [10].

Definition 4.0.1. The *generalized linear model* is a statistical model with these components:

1. *The random component* - The response variables $Y_1, Y_2, \ldots, Y_n$ are independent identically distributed with distribution from the exponential family in canonical form, thus we have

   $$f(y_i; \theta_i) = \exp\left[y_i b(\theta_i) + c(\theta_i) + d(y_i)\right]$$

   and the joint pdf

   $$\begin{aligned} f(y_1, y_2, \ldots, y_n; \theta_1, \theta_2, \ldots, \theta_n) &= \prod_{i=1}^{n} \exp\left[y_i b(\theta_i) + c(\theta_i) + d(y_i)\right] \\ &= \exp\left[\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right]. \end{aligned}$$

2. *The systematic component* - The vector $(\eta_1, \ldots, \eta_n)$ relates to the explanatory variables through a linear model. Let x_{ij} denote the value of predictor $j (j = 1, \ldots, p)$ for subject i. Then the *linear predictor* is

   $$\eta_i = \sum_j \beta_j x_{ij}, \quad i = 1, \ldots, n.$$

3. *The link function* - Suppose that for some functions μ_i of θ_i the expected value is $\mathrm{E}[Y_i] = \mu_i, i = 1, \ldots, n$. The model links μ_i to η_i by $\eta_i = g(\mu_i)$. The function g is the link function, it is monotone and differentiable.

Note 4.0.1. We can also write the link function in vector form as $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$. Then the column vector of explanatory variables, which is the i-th column of the design matrix, and column vector of parameters are as follows

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ & \ddots & \\ x_{n1} & \ldots & x_{np} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

4.1 Score and Fisher information

Let us have a GLM specified in definition 4.0.1. For each variable Y_i we have the log-likelihood function, expected value, variance, and link function as

$$\begin{aligned} l_i(\theta_i) &= y_i b(\theta_i) + c(\theta_i) + d(y_i) \\ \mathrm{E}[Y_i] &= \mu_i = -\frac{c'(\theta_i)}{b'(\theta_i)} \\ \mathrm{var}[Y_i] &= \frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{(b'(\theta_i))^3} \\ g(\mu_i) &= \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i. \end{aligned}$$

The log-likelihood function for all the Y_i's is

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i).$$

By differentiating with respect to β_j we obtain the score

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \tag{4.1.1}$$

using the chain rule for differentiations. Let us consider the terms separately. We can assign the first term as

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = y_i b'(\theta_i) - \mu_i b'(\theta_i) = b'(\theta_i)(y_i - \mu_i).$$

The middle term can be reassessed as

$$\frac{\partial \theta_i}{\partial \mu_i} = 1/\left(\frac{\partial \mu_i}{\partial \theta_i}\right) = 1/\left(\frac{-c''(\theta_i)b'(\theta_i) + c'(\theta_i)b''(\theta_i)}{(b'(\theta_i))^2}\right) = \frac{1}{b'(\theta_i)\text{var}\left[Y_i\right]}.$$

The last term is

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}.$$

Then the score is given by

$$U_j = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\text{var}\left[Y_i\right]}\right) x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right). \tag{4.1.2}$$

The variance-covariance matrix of U_j's has terms $\mathfrak{J}_{jk} = \text{E}\left[U_j U_k\right]$ which forms the Fisher information matrix $\mathfrak{J}$. From 4.1.2 we obtain

$$\begin{aligned} \mathfrak{J}_{jk} &= \text{E}\left[\sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\text{var}\left[Y_i\right]}\right) x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) \cdot \sum_{l=1}^{n} \left(\frac{Y_l - \mu_l}{\text{var}\left[Y_l\right]}\right) x_{lk} \left(\frac{\partial \mu_l}{\partial \eta_l}\right)\right] \\ &= \text{E}\left[\sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\text{var}\left[Y_i\right]}\right)^2 x_{ij} x_{ik} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right], \end{aligned} \tag{4.1.3}$$

as $\text{E}\left[(Y_i - \mu_i)(Y_l - \mu_l)\right] = 0$ for $i \neq l$ because the Y_i's are independent. Then using the fact that $\text{E}\left[(Y_i - \mu_i)^2\right] = \text{var}\left[Y_i\right]$ we obtain

$$\mathfrak{J}_{jk} = \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{\text{var}\left[Y_i\right]} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \tag{4.1.4}$$

which can be written as

$$\mathfrak{J} = \mathbf{X}^T \mathbf{W} \mathbf{X},$$

where $\mathbf{W}$ is a $N \times N$ diagonal matrix given by

$$w_{ii} = \frac{1}{\text{var}\left[Y_i\right]} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \quad i = 1, \ldots, n. \tag{4.1.5}$$

4.2 Newton - Raphson method

Newton - Raphson method is a numerical iterative procedure which is used to solve nonlinear equations, such as finding a maximum of a function. Therefore it allows us to find the estimators $\hat{\boldsymbol{\beta}}$ for the GLM. It starts with an initial

estimator for the solution and by iteration it approximates the function to be maximized in a neighbourhood of the previous estimation by a second-degree polynomial and it obtains an estimated location of the maximum. This sequence of estimated locations converges to the actual location of the maximum, if the initial estimate is good enough.

Let $\mathbf{u}' = \left(\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_p}\right)$ be the gradient of a function $L(\boldsymbol{\beta})$, let the *Hessian matrix* $\mathbb{H}$

$$\mathbb{H} = \begin{bmatrix} \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1^2} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p} \\ \vdots & & \vdots \\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_p^2} \end{bmatrix}.$$

be regular. Let $\mathbf{u}^{(t)}$ and $\mathbb{H}^{(t)}$ be $\mathbf{u}$ and $\mathbb{H}$ evaluated at $\boldsymbol{\beta}^{(t)}$ as t-th estimate for $\boldsymbol{\beta}$. By Tailor series expansion we can approximate the likelihood function as

$$L(\boldsymbol{\beta}) \approx L(\boldsymbol{\beta}^{(t)}) + \mathbf{u}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^T \mathbb{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}).$$

Solving the equation $\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx \mathbf{u}^{(t)} + \mathbb{H}^{(t)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) = 0$ for $\boldsymbol{\beta}$ yields the next guess

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\mathbb{H}^{(t)})^{-1}\mathbf{u}^{(t)}.$$

The convergence of $\boldsymbol{\beta}^{(t)}$ to $\hat{\boldsymbol{\beta}}$ for large t satisfies

$$|\beta_j^{(t+1)} - \hat{\beta}_j| \leq c|\beta_j^{(t)} - \hat{\beta}_j|^2$$

for $j = 1, \ldots, p$ and for some $c > 0$.

Note 4.2.1. In GLM modified Newton-Raphson method is used. It leads to iterations obtained by solving the system of equations

$$\mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{X}\boldsymbol{\beta}^{(m)} = \mathbf{X}^T\mathbf{W}^{(m-1)}\mathbf{Z}^{(m-1)},$$

where $\mathbf{W}^{(m-1)}$ the value of matrix $\mathbf{W}$, defined in the previous section, at m-th iteration and $Z_i = \sum_{k=1}^{p} x_{ik}\beta_k^{(m-1)} + (Y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$. For more see [6].

4.3 Goodness of fit

Models of n observations can be fitted containing up to n parameters. The simplest model is the null model, concerning just one parameter, which represents the common mean for all y's. Such model distributes all the variation between the y's to the random component. On the other hand, the full

model has n parameters and it distributes all the variation in the y's to the systematic component leaving none for the random component.

Usually none of these two extreme models is used in practice as the null model is too simple and the full model doesn't leave any randomness. However, the full model is a good baseline for measuring the discrepancy for an intermediate model with p parameters.

Definition 4.3.1. Let us have a studied model with $\hat{\boldsymbol{\theta}}$ its ML parameter estimate and full model with $\hat{\boldsymbol{\theta}}_{max}$ its ML parameter estimate. Then the discrepancy

$$D = 2l(\hat{\boldsymbol{\theta}}_{max}; \mathbf{Y}) - 2l(\hat{\boldsymbol{\theta}}, \mathbf{Y}),$$

is called the *scaled deviance* for the studied model.

Another approach is using a criterion, which judges the model by how close its fitted values tend to be to the true values, in terms of a certain expected value. It takes into account both the statistical goodness of fit and the number of parameters used in a model, and imposes a penalty for increasing this number.

Definition 4.3.2. Let us select a model, where $\hat{\boldsymbol{\theta}}$ is the estimate of parameter vector $\boldsymbol{\theta}$ and p the number of parameters for this model. Then we want to minimize the formula

$$AIC = -2(l(\hat{\boldsymbol{\theta}}; \mathbf{Y}) - p)$$

which we call the *Akaike information criterion*.

4.4 Residuals

Residuals can be used as a tool to explore the adequacy of a fit of a model. They may indicate any anomalies and outliers that would need further investigation. For these purposes we have to replace the usual standard Normal residuals with an extended definition of residuals, i.e. generalized residuals.

Definition 4.4.1. We define the *Pearson residual* as

$$r_P = \frac{y - \mu}{\sqrt{V(\mu)}},$$

where $\sqrt{V(\mu)}$ is the standard deviation and μ is the expected value of a random variable Y.

Note 4.4.1. The Pearson residual is the standard residual $y - \mu$ scaled by the standard deviation. These residuals don't fit situations, when the distribution of the r_P's is considerably skewed and therefore it fails to have properties similar to those of standard Normal residuals.

In order to describe the residuals feasibly Francis J. Anscombe proposed a function $A(y)$ to be used instead of y making the distribution of $A(Y)$ 'as Normal as possible'.

Definition 4.4.2. Let the function

$$A(\mu) = \int_{-\infty}^{\mu} \frac{dt}{\sqrt[3]{V(t)}}.$$

Then the *Anscombe residual* is the

$$r_A = \frac{A(y_i) - A(\mu_i)}{A'(\mu_i)\sqrt{V(\mu_i)}},$$

where $A'(\mu_i)$ is the derivative of $A(\mu_i)$ with respect to μ_i.

Example 4.4.1. We name Anscombe residuals for several distributions according to [10].

(a) Poisson distribution

$$r_A = \frac{\frac{3}{2}(y^{2/3} - \mu^{2/3})}{\mu^{1/3}},$$

(b) Binomial distribution

$$r_A = \sqrt{n_i}\frac{B(y_i, \frac{2}{3}, \frac{2}{3}) - B(\mu_i, \frac{2}{3}, \frac{2}{3})}{\frac{\mu_i}{1-\mu_i}},$$

where n_i is the number of trials in i-th trial and

$$B(z, a, b) = \int_0^z t^{a-1}(1-t)^{b-1}dt,$$

(c) Negative Binomial distribution

$$r_A = \frac{\frac{3}{2}y^{2/3}F(\frac{1}{3}, \frac{2}{3}, \frac{5}{3}, -\frac{y}{\kappa}) - \frac{3}{2}\mu^{2/3}F(\frac{1}{3}, \frac{2}{3}, \frac{5}{3}, -\frac{\mu}{\kappa})}{(\mu + \frac{\mu^2}{\kappa})^{1/6}},$$

where F is the hypergeometric function.

Definition 4.4.3. Let D be the deviance and d_i part of the deviance in i-th unit such that $D = \sum_i d_i$. Then the *deviance residual* is

$$r_D = \text{sign}(y_i - \mu_i)\sqrt{d_i}.$$

5 Quasilikelihood methods

As seen from examples, some of the distributions were not suitable for the ML method. Therefore we introduce the quasilikelihood method, which takes into consideration the relationship between expected value and variance of the response variables. Main source for this chapter was [10].

5.1 Quasi-likelihood function

Let us suppose that the components of the response vector $\mathbf{Y}$ are independent, they have a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\sigma^2\mathbf{V}(\boldsymbol{\mu})$, where σ^2 may be unknown and $\mathbf{V}(\boldsymbol{\mu})$ is a matrix of known functions. We will assume that the vector of parameters $\boldsymbol{\beta}$ relate to the dependence of $\boldsymbol{\mu}$ on covariates $\mathbf{x}$, therefore we can write $\boldsymbol{\mu}(\boldsymbol{\beta})$. Also we assume that σ^2 does not depend on β.

From the assumption of independence of the components of $\mathbf{Y}$ we can write the matrix as $\mathbf{V}(\boldsymbol{\mu}) = diag\left\{V_1(\boldsymbol{\mu}), \ldots, V_n(\boldsymbol{\mu})\right\}$. Furthermore the function $V_i(\boldsymbol{\mu})$ depends only on the i-th component of $\boldsymbol{\mu}$, thus we can write

$$\mathbf{V}(\boldsymbol{\mu}) = diag\left\{V_1(\mu_1), \ldots, V_n(\mu_n)\right\}. \tag{5.1.1}$$

Now consider only a single component Y of the response vector $\mathbf{Y}$ and a function

$$U = u(\boldsymbol{\mu}; Y) = \frac{Y - \boldsymbol{\mu}}{\sigma^2 V(\boldsymbol{\mu})},$$

which has similar properties as the log-likelihood derivative

$$\begin{aligned} E\left[U\right] &= 0 \\ var\left[U\right] &= \frac{1}{\sigma^2 V(\boldsymbol{\mu})} \\ -E\left[\frac{\partial U}{\partial \boldsymbol{\mu}}\right] &= \frac{1}{\sigma^2 V(\boldsymbol{\mu})}. \end{aligned}$$

The integral

$$Q(\boldsymbol{\mu}; y) = \int_y^{\boldsymbol{\mu}} \frac{y-t}{\sigma^2 V(t)} dt,$$

if it exists, behaves like a log-likelihood function for $\boldsymbol{\mu}$ and is called *log quasi-likelihood* for $\boldsymbol{\mu}$ based on data y. The quasi-likelihood function for the complete data is the sum of the individual contributions

$$Q(\boldsymbol{\mu}, \mathbf{y}) = \sum Q_i(\mu_i, y_i),$$

as the components of the random vector Y are independent.

By analogy, the quasi-deviance function corresponding to a single observation is

$$D(y; \boldsymbol{\mu}) = -2\sigma^2 Q(\boldsymbol{\mu}; y) = 2 \int_y^{\boldsymbol{\mu}} \frac{y-t}{V(t)} dt. \tag{5.1.2}$$

5.2 Parameter estimation

We use similar approach to finding the estimators β, we differentiate the quasi-likelihood function $Q(\boldsymbol{\mu}, \mathbf{y})$ with the respect to β

$$\mathbf{U}(\beta) = \mathbf{D^T V^{-1}} (\mathbf{Y} - \boldsymbol{\mu})/\sigma^2; \tag{5.2.1}$$

where the matrix $\mathbf{D}$ of order $n \times p$ are the derivatives of $\boldsymbol{\mu}(\beta)$ with the respect to the parameters β, therefore the components of such matrix are $D_{ij} = \partial \boldsymbol{\mu}_i / \partial \beta_j$. The function $\mathbf{U}(\beta)$ is called the *quasi-score function*.

The covariance matrix of $\mathbf{U}(\beta)$ being a negative expected value of $\partial \mathbf{U}/\partial \beta$ is

$$\mathbf{i}_\beta = \mathbf{D^T V^{-1} D}/\sigma^2, \tag{5.2.2}$$

which plays the same role as Fisher information for ordinary likelihood functions.

If we begin with an arbitrary value $\widehat{\beta_\mathbf{0}}$ sufficiently close to the estimator $\widehat{\beta}$, the sequence of parameter estimates given by the Newton-Raphson method with Fisher scoring is

$$\widehat{\beta_\mathbf{1}} = \hat{\beta}_0 + (\hat{D}_0^T \hat{V}_0^{-1} \hat{D}_0)^{-1} \hat{D}_0^T \hat{V}_0^{-1} (y - \hat{\boldsymbol{\mu}}_0).$$

By iterating until the convergence occurs we will obtain the quasi-likelihood estimate $\hat{\beta}$. Such sequence does not depend on the value of σ^2.

Under the usual conditions on the eigenvalues of $\mathbf{i}_\beta$, the asymptotic covariance matrix of $\widehat{\beta}$ is given by

$$cov(\widehat{\beta}) \backsimeq \mathbf{i}_\beta^{-\mathbf{1}} = \sigma^2 \mathbf{D^T V^{-1} D^-} 1.$$

6 Application on real data

The data we are about to analyze resemble the nature of random sums, i.e. the sum of independent identically distributed random variables, where the number of summands is also a random variable. Firstly, we will discuss overall graphical analysis of the data, then we will compare the graphical attributes of our data with the computed ones, then we will discuss the generalized linear models and quasi-likelihood method that could correspond to our case taking in consideration different distributions as well. The whole analysis was made by programming language R of version 3.4.4.

6.1 Graphical analysis

For the graphical part of analysis we were using boxplots and histograms of the variables Bay, $Block$, $Position$, $Split$, $Treated$, and $Treatment$. Here we show just some of the results, the rest can be found in the attachments.

6.1.1 Moorlands

From the histogram of the data, we cannot assess the shape of the distribution very precisely. From Figure 6.1 it appears to have two local maxima for around 20 and 40. The rest of the histograms does not give us more information about the distribution.

For the boxplots by Treatment 6.2 we see that for the treatments $T26$, $T27$, $T28$, and $T30$ the data are skewed right with some outliers between 40 and 60. Unfortunately, for treatments $29TD$, $42TD$, $62TD$, and $6TD$ there were only 10 samples, therefore we cannot make any strong conclusions from their boxplots.

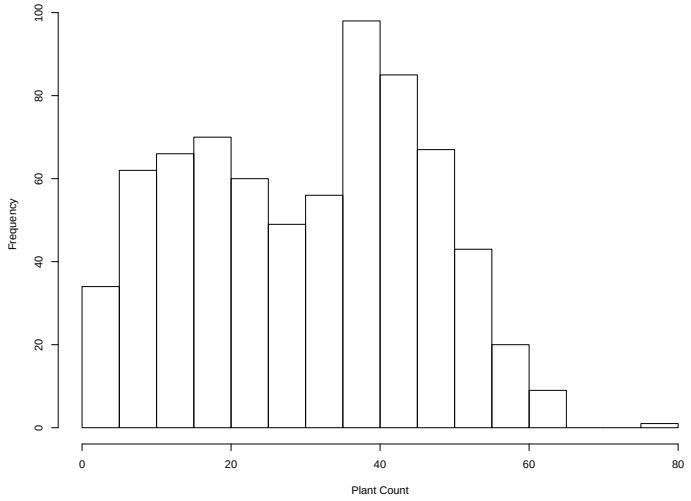


Figure 6.1: Moorlands - histogram of Plant Count

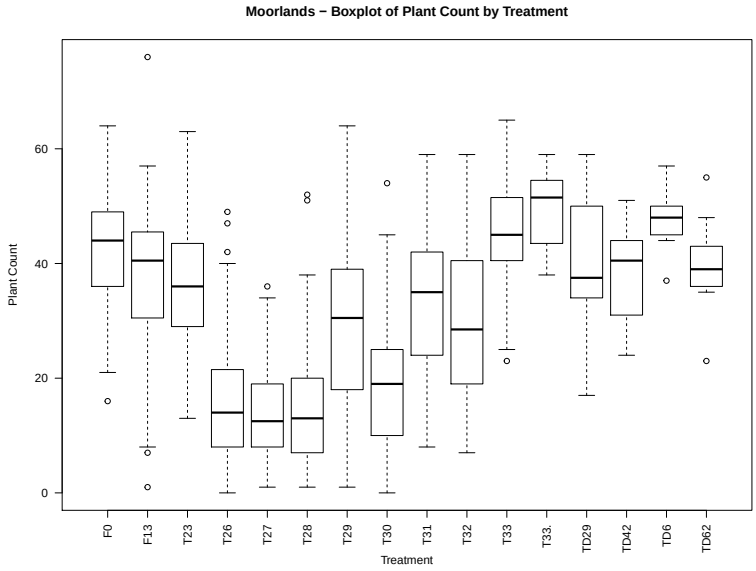
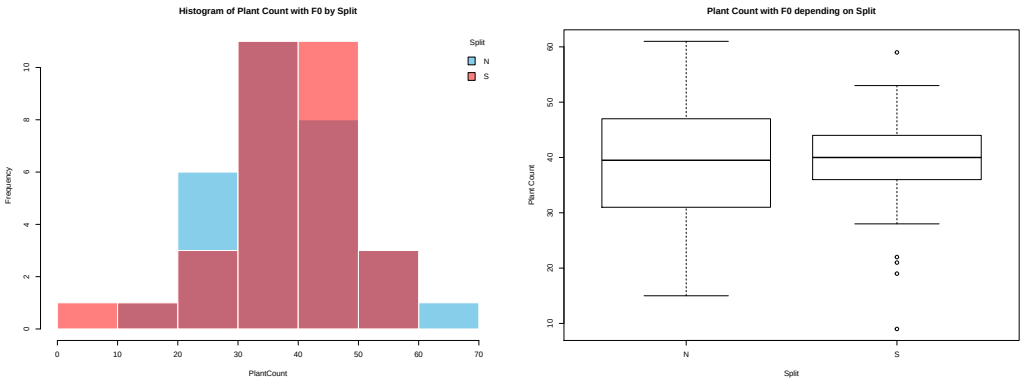


Figure 6.2: Moorlands - boxplot of Plant Count by Treatment

There was a slight concern that the left and right side of the seeder could behave differently, mainly confounded with the variable *Split*. Therefore the section with $Treatment = F0$ was implemented for the comparison without effect of the $Treatment$. The difference in the sides of the seeder can be assessed by viewing Figure 6.3.



(a) Histogram (b) Boxplot

Figure 6.3: Moorlands - Graphs for the Treatment F0 by Split

6.1.2 Geranium

In comparison with Moorlands, Geranium produces histogram nearly symmetrical with the modus around 40, which can be seen in Figure 6.4. According to the boxplot in Figure 6.5 the number of *PlantCount* is higher than for Moorlands, the median is very similar for each boxplot and it is between 30 and 40.

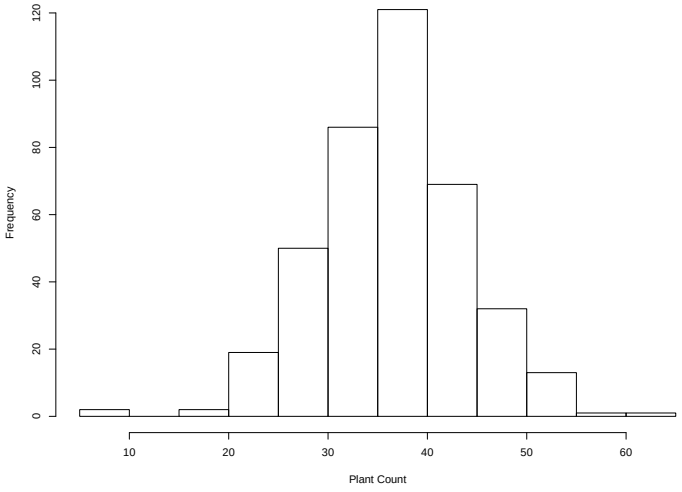


Figure 6.4: Geranium - histogram of Plant Count

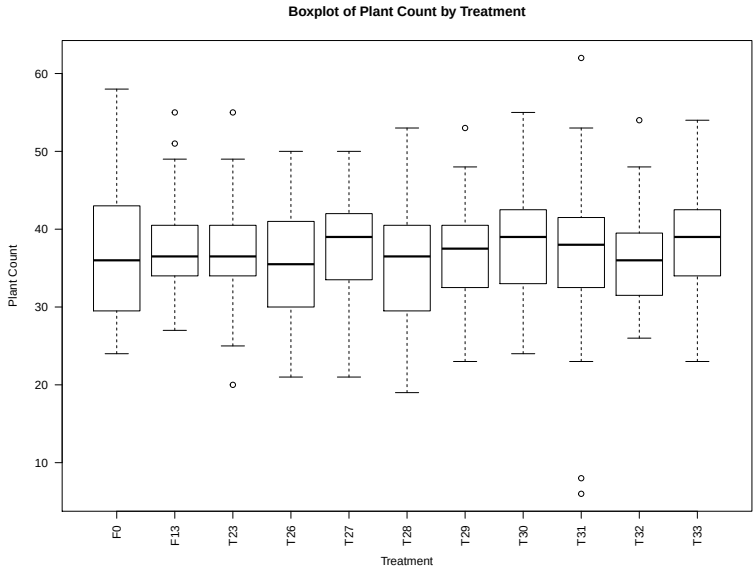


Figure 6.5: Geranium - boxplot of Plant Count by Treatment

6.2 Mean - variance relationship

Having previously determined the relationship between expected value and variance for different random sums in section 2.1, we can compare the results with our data.

For both areas we computed the mean and variance by the variable *Plot* and then we divided them by the combinations of $Treatment$ or *Block* and *Split* and $Treated$. These figures can be found in the Appendix.

Unfortunately, our expectations weren't met and we weren't able to find any relationship that would correlate with our assumptions. There is no evidence of a common trend in the graphs and even though the data comported with the idea of random sums, we couldn't confirm our hypothesis.

6.3 Generalized Linear Models

As was derived in the Chapter 4, three of the considered distributions are of the exponential family and therefore can be analyzed by GLM. We are using the function `glm` already implemented in R. It takes at least two input parameters, the formula and the family of the error distribution with link function.

The model we take into consideration is the formula

$$PlantCount \sim Treatment * Treated * Split * Block * Bay, \qquad (6.3.1)$$

where $PlantCount$ is the response variable and $Treatment$, $Treated$, $Split$, $Block$, and Bay are the explanatory variables. Let us have two random variables $V1, V2$, then the notation $V_1 * V_2$ is equal to $\beta_1 V_1 + \beta_2 V_2 + \beta_3 V_1 \cdot V_2$, it covers both the summation and the interaction of the variables.

For each model we obtain two deviances, null and residual. The null deviance shows how well does the model including only the intercept, i.e. the null model, predict the response variable, whereas the residual deviance tells us how well is the response variable predicted by proposed model. The degrees of freedom for each deviance is the number of observations subtracted by the number of predictors.

Then the function `step` can be used to choose a model by AIC in a stepwise algorithm, or we can compare the models with the function `anova` by test based on deviance (in `R` denoted by `Chisq`) and then using the functions `drop1` and `update` we identify the most suitable model.

After finding the most suitable model for each distribution, we determine the Anscombe residuals according to the used model.

In the binomial models the response variable needs to be in the interval $\langle 0, 1 \rangle$, therefore we need to find appropriate parameter n, which would divide the variable. For this particular case we estimated the parameter by Jackknife estimator [5].

6.3.1 Moorlands

Binomial distribution

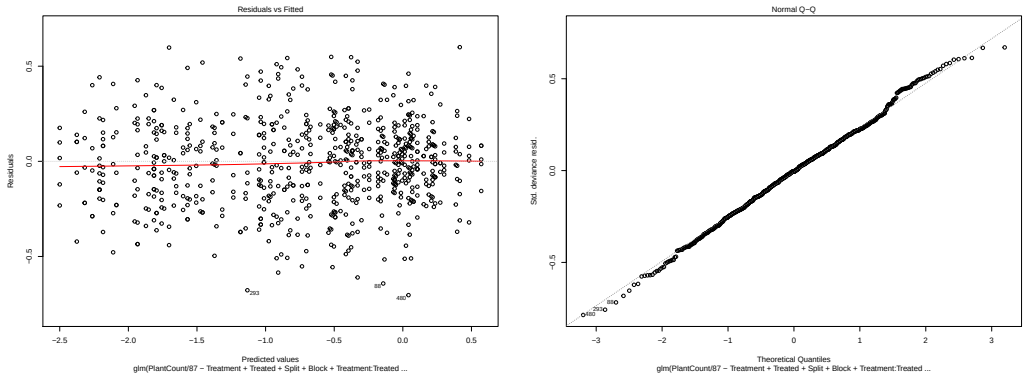
As the divisor we have estimated the parameter $n = 87$. By dropping insignificant variables we get a model

$$\begin{aligned} PlantCount/87 \sim\ & Treatment + Treated + Split + Block \\ & + Treatment : Treated + Treatment : Bay + Treated : Split \\ & + Treated : Block + Split : Block + Treatment : Treated : Bay \\ & + Treatment : Split : Block : Bay. \end{aligned} \qquad (6.3.2)$$

Table 6.1: Moorlands: Binomial family

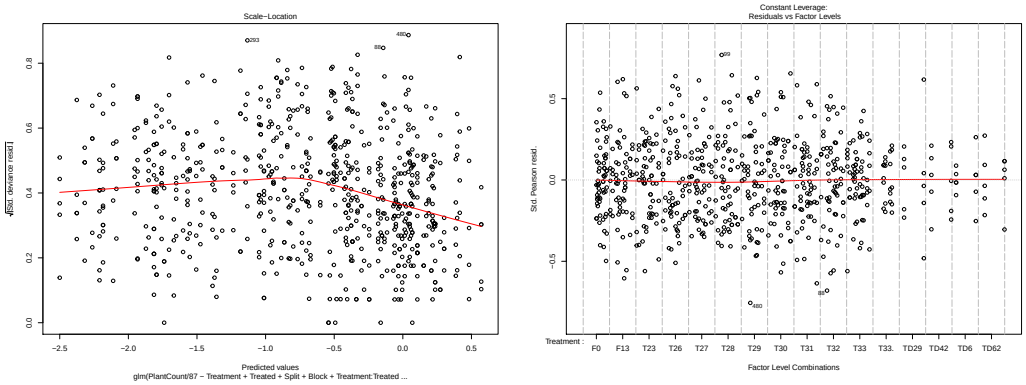
	Null deviance	Residual deviance
Selected model	114.580 on 719 df	35.885 on 576 df

As can be seen from Table 6.1 and Figure 6.6 the model describes the data well.



(a) Residuals vs. Fitted values

(b) Normal Q-Q plot



(c) Scale - Location graph

(d) Residuals vs Factor Levels graph

Figure 6.6: Moorland - Graphical analysis of GLM with Binomial family

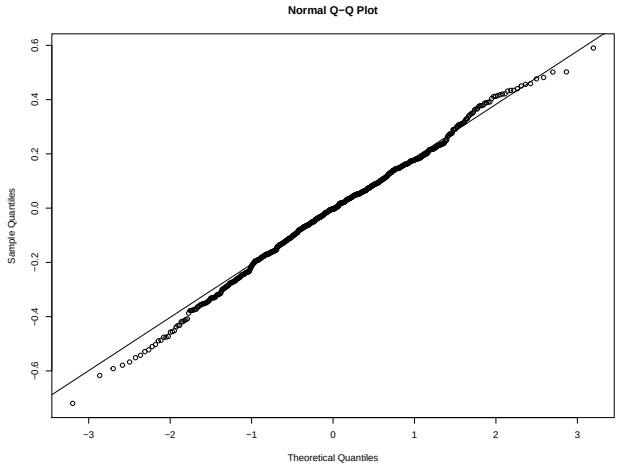


Figure 6.7: Moorlands - Anscombe residuals for Binomial family

Poisson distribution

Next option to choose was the Poisson distribution. Dropping the variable *Bay* and updating the initial model 6.3.1 we get the formula

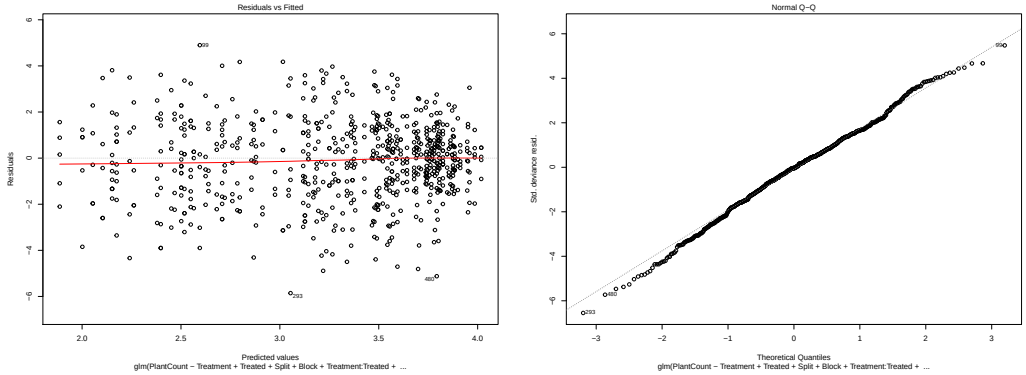
$$\begin{aligned} PlantCount \sim\ & Treatment + Treated + Split + Block \\ & + Treatment : Treated + Treatment : Block + Treated : Block \\ & + Split : Block + Treatment : Treated : Block, \end{aligned}$$

Results can be observed in Table 6.2.

Table 6.2: Moorlands: Poisson family

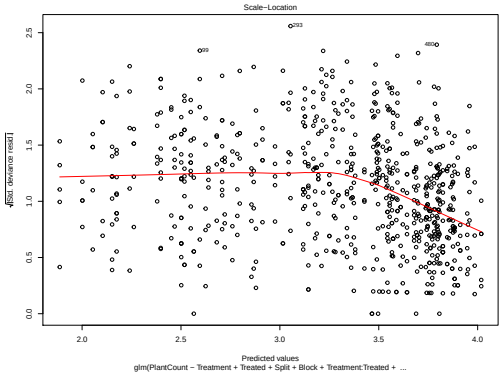
	Null deviance	Residual deviance
Selected without Bay	6688.3 on 719 df	2104.6 on 576 df

The presented model is the best for Poisson family but doesn't fit the data well as we can see from Table 6.2 and from Anscombe residuals in Figure 6.9.

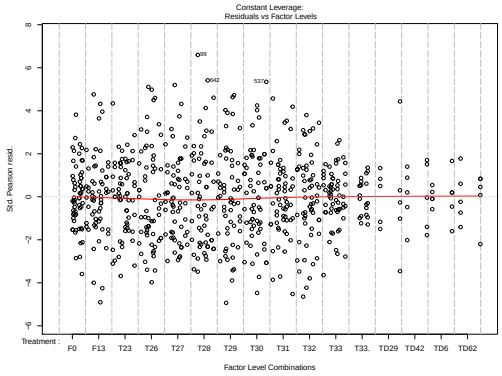


(a) Residuals vs. Fitted values

(b) Normal Q-Q plot



(c) Scale - Location graph



(d) Residuals vs Factor Levels graph

Figure 6.8: Moorland - Graphical analysis of GLM with Poisson family

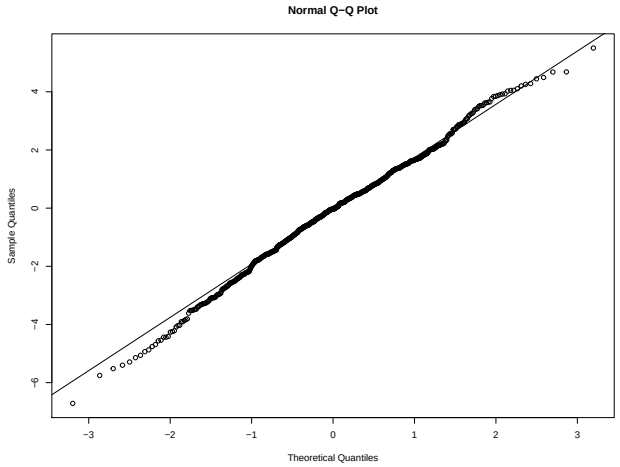


Figure 6.9: Moorlands - Anscombe residuals for Poisson family

Negative binomial distribution

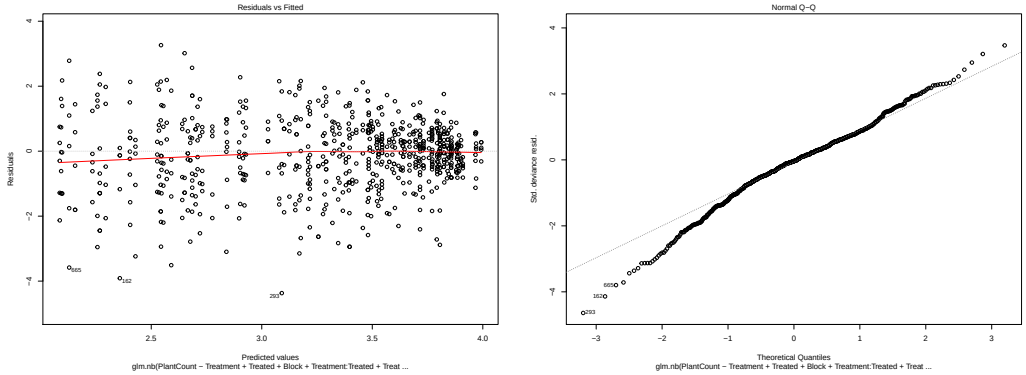
For the negative binomial family we used the library `MASS` with the function `glm.nb`. By altering the initial model we obtain

$$PlantCount \sim Treatment + Treated + Block + Treatment : Treated + Treatment : Block.$$

As can be observed from Table 6.3 and from Figure 6.10, especially from the Q-Q plot, this choice of family doesn't suit our data well.

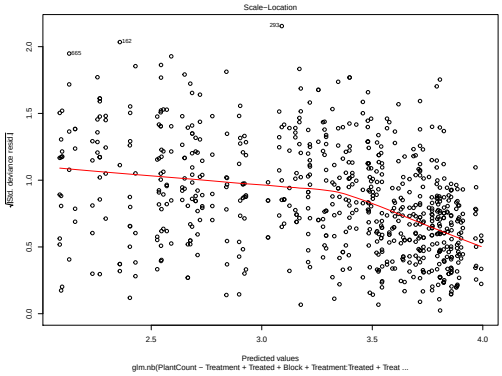
Table 6.3: Moorlands: Negative binomial family

	Null deviance	Residual deviance
Selected model	2261.87 on 719 df	839.03 on 633 df

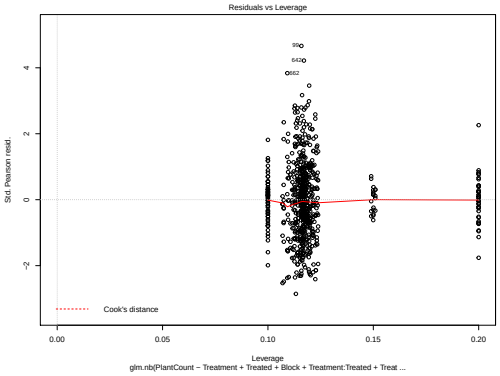


(a) Residuals vs. Fitted values

(b) Normal Q-Q plot



(c) Scale - Location graph



(d) Residuals vs Factor Levels graph

Figure 6.10: Moorland - Graphical analysis of GLM with Negative binomial family

6.3.2 Geranium

For most of the models we dropped the variable *Bay*, because it might be bound to the variable *Block*. Several outliers can be observed in all models, namely points $121, 122, 124$. All three of them had $Bay = 1$, $Block = 2$, $Treatment = T31$ and $Plot = 21$, other properties are in Table 6.4.

Table 6.4: Geranium: Outliers

Number	Split	Treated	PlantCount
121	S	Treated	6
122	S	Treated	62
124	N	Untreated	8

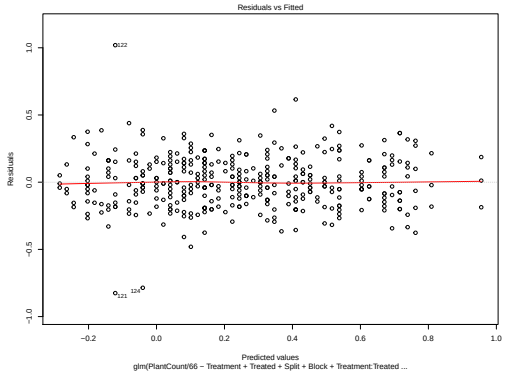
Binomial distribution

We estimated the parameter n to be 66. Using the initial model and altering it by dropping the variable *Bay* produces same results, which we present in Table 6.5. By applying the function `step` to the initial model we receive the null model containing only the intercept, which doesn't describe the data at all as it is not representing the randomness. From manually updating we obtained

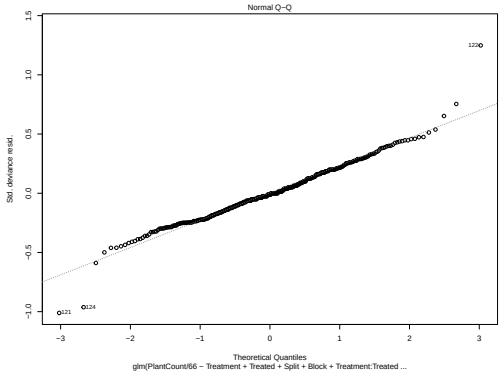
$$\begin{aligned} PlantCount/66 \sim\ & Treatment + Treated + Split + Block \\ & + Treatment : Treated + Treatment : Split + Treatment : Block \\ & + Treated : Block + Treatment : Treated : Block. \end{aligned} \quad (6.3.3)$$

Table 6.5: Geranium: Binomial family

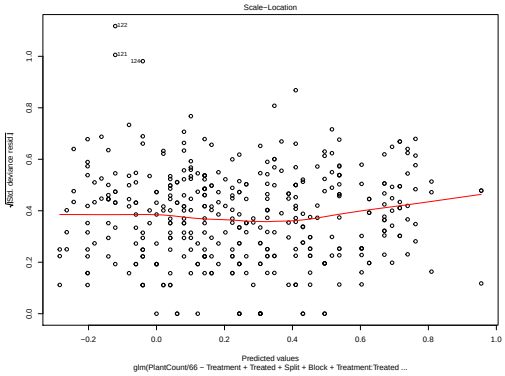
	Null deviance	Residual deviance
Selected model	21.649 on 395 df	14.267 on 264 df



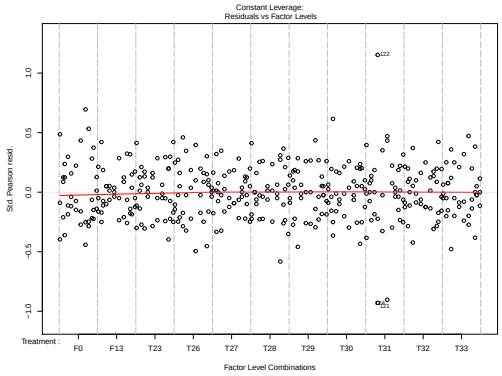
(a) Residuals vs. Fitted values



(b) Normal Q-Q plot



(c) Scale - Location graph



(d) Residuals vs Factor Levels graph

Figure 6.11: Geranium - Graphical analysis of GLM with Binomial family

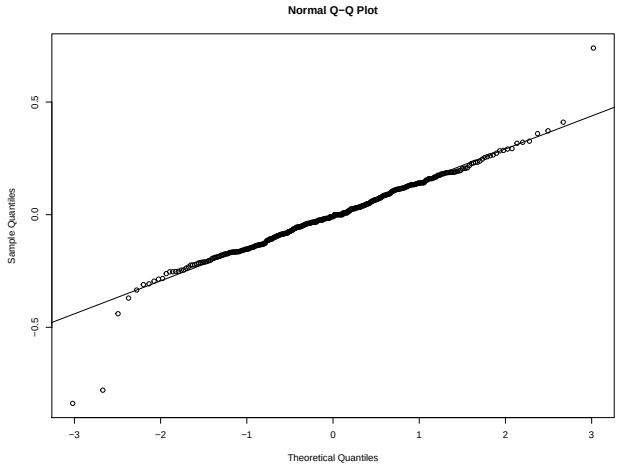


Figure 6.12: Geranium - Anscombe residuals for Binomial family

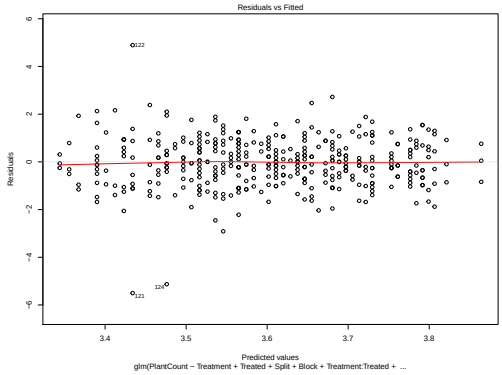
Poisson distribution

By altering the initial model we get a formula

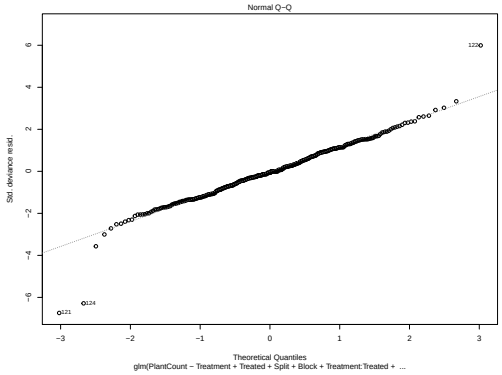
$$\begin{aligned} PlantCount \sim\ & Treatment + Treated + Split + Block \\ & + Treatment : Treated + Treatment : Split + Treatment : Block \\ & + Treated : Block + Treatment : Treated : Block. \end{aligned}$$

Table 6.6: Geranium: Poisson family

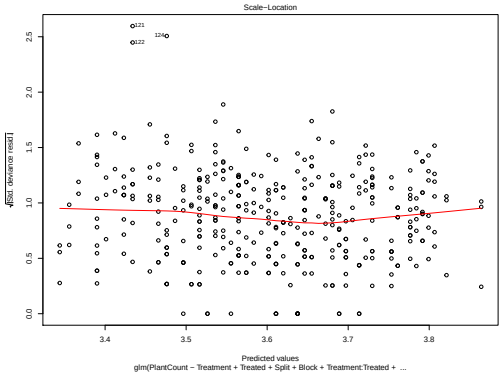
	Null deviance	Residual deviance
Selected model	627.59 on 395 df	415.64 on 264 df



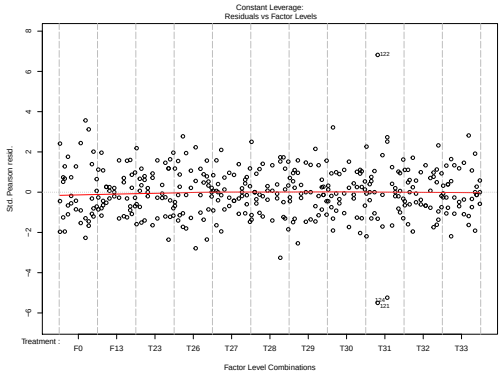
(a) Residuals vs. Fitted values



(b) Normal Q-Q plot



(c) Scale - Location graph



(d) Residuals vs Factor Levels graph

Figure 6.13: Geranium - Graphical analysis of GLM with Poisson family

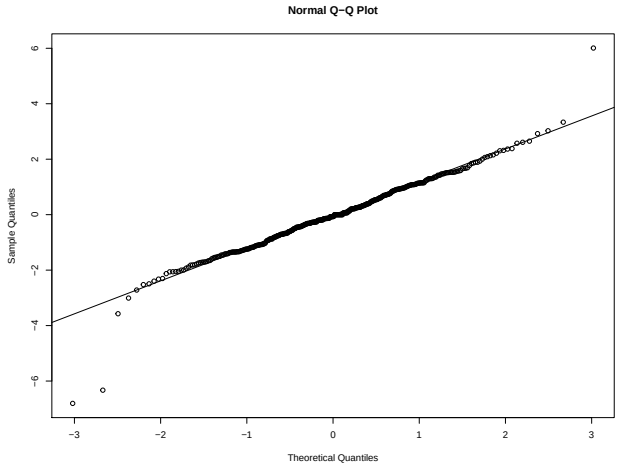


Figure 6.14: Geranium - Anscombe residuals for Poisson family

Negative binomial distribution

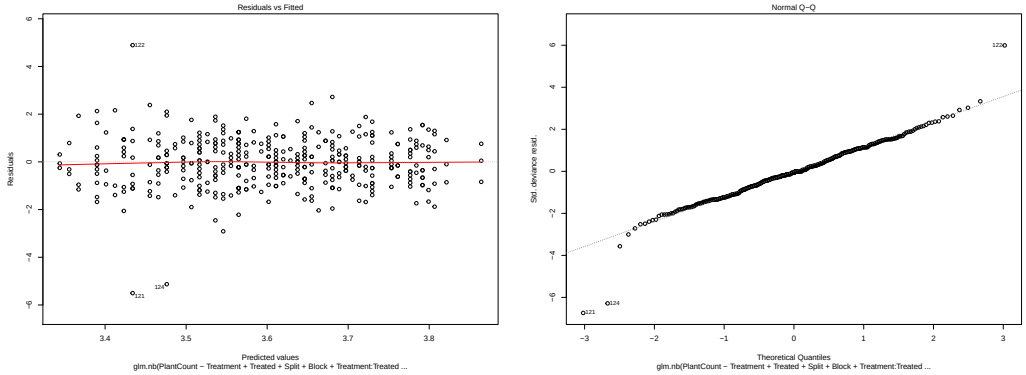
By omitting the variable *Bay* and updating the initial model we got

$$\begin{aligned} PlantCount \sim Treatment + Treated + Split + Block + Treatment:Treated \\ + Treatment:Split + Treated:Split \\ + Treatment:Block + Treated:Block + Split:Block \\ + Treatment:Treated:Split + Treatment:Treated:Block, \end{aligned}$$

which is much simpler. The results can be seen in Table 6.7.

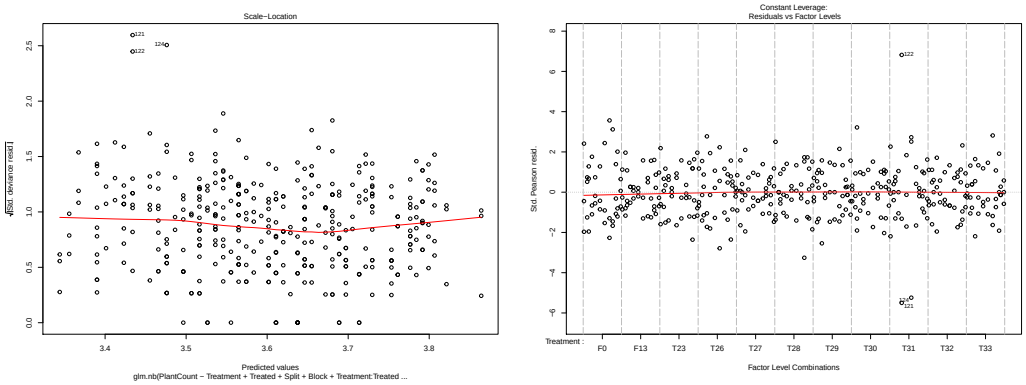
Table 6.7: Geranium: Negative binomial family

	Null deviance	Residual deviance
Selected model	627.47 on 395 df	415.57 on 264 df



(a) Residuals vs. Fitted values

(b) Normal Q-Q plot



(c) Scale - Location graph

(d) Residuals vs Factor Levels graph

Figure 6.15: Geranium - Graphical analysis of GLM with Negative binomial family

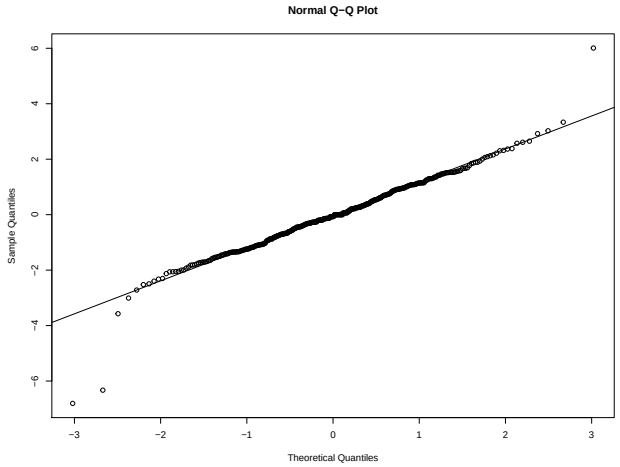


Figure 6.16: Geranium - Anscombe residuals for Negative binomial family

6.4 Quasi Models

From the Chapter 4 we observed that the mixtures of discrete uniform-Bernoulli and logarithmic-Bernoulli distributions don't have any known distribution and they are not of the exponential family, hence they cannot be analyzed by GLM. Therefore we are using the quasi family in the function `glm` with different link function and variance that would suit our problem. As the variance we will use previously determined relationship of expected value and variance from 2.1.

6.4.1 Moorlands

Discrete uniform - Bernoulli distribution

Let us recall relationship between the expected value and variance for distribution $U(a,b) - Be(p)$ as follows

$$V(\mu) = \mu^2 \left(-\frac{2}{a+b} + \frac{(b-a)(b-a+2)}{3(a+b)^2} \right) + \mu,$$

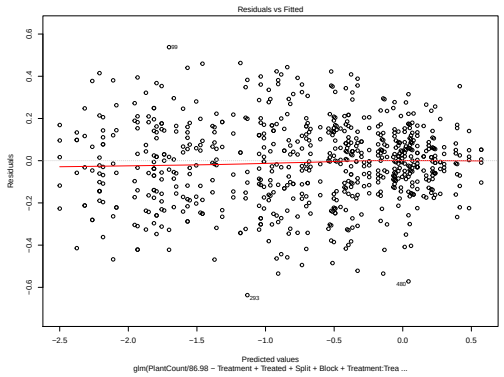
which is going to be our variance and the link function is `logit`. We have chosen $a = 0$ and $b = 100$ and by altering the initial model we obtained

$$\begin{aligned} PlantCount/87 \sim\ & Treatment + Treated + Split + Block \\ & + Treatment : Treated + Treatment : Split + Treatment : Block \\ & + Treated : Block + Split : Block + Treatment : Treated : Block. \end{aligned}$$

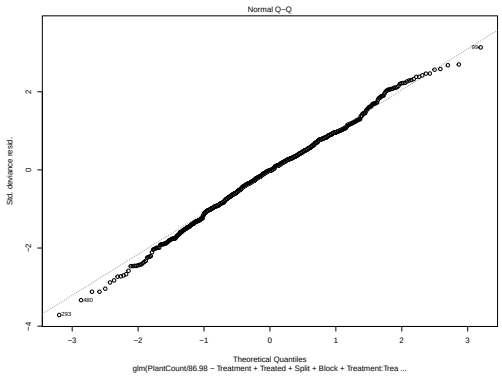
The dispersion parameter σ^2 estimated in the model was 0.04293782. Even though the residual deviance is very small, from the graphical analysis we can see that this model is not suitable.

Table 6.8: Moorlands: Quasi with U-Be

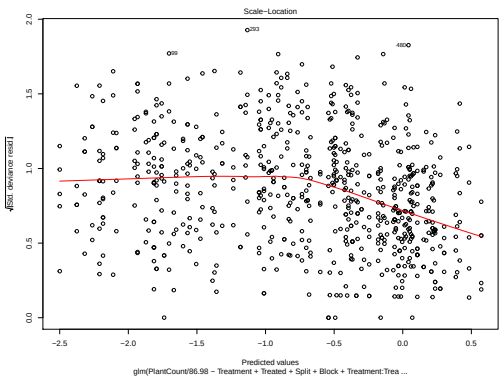
	Null deviance	Residual deviance
Selected model	82.113 on 719 df	25.738 on 576 df



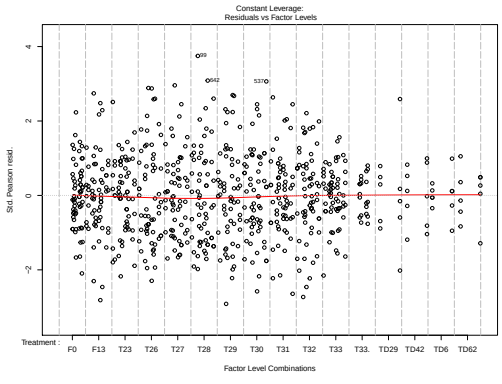
(a) Residuals vs. Fitted values



(b) Normal Q-Q plot



(c) Scale - Location graph



(d) Residuals vs Factor Levels graph

Figure 6.17: Moorlands - Graphical analysis of quasi GLM with U - Be

Logarithmic - Bernoulli distribution

Let us recall relationship between the expected value and variance for distribution $Log(q) - Be(p)$ as follows

$$V(\mu) = \mu\left(1 - \mu(\ln\left(1-q\right)+1)\right),$$

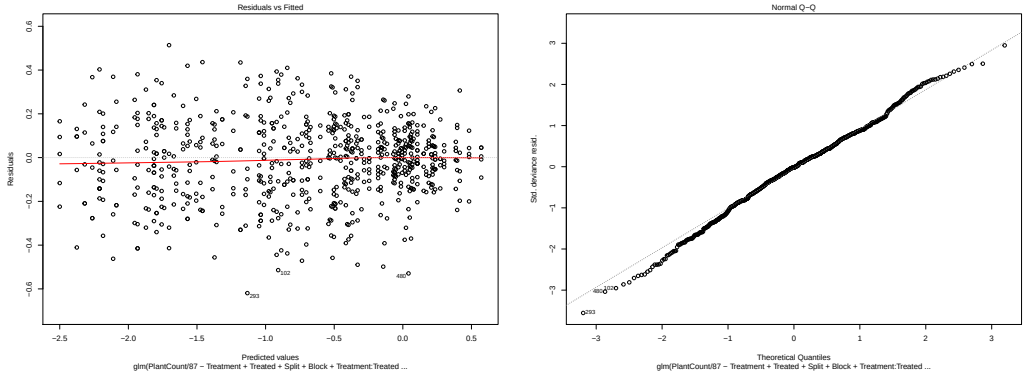
which is going to be our variance and the link function is `logit`. We have chosen $q = 0.7$ and by altering the initial model we obtained

$$\begin{aligned} PlantCount/87 \sim\ & Treatment + Treated + Split + Block \\ & + Treatment : Treated + Treatment : Split + Treatment : Block \\ & + Treated : Block + Split : Block + Treatment : Treated : Block. \end{aligned}$$

The dispersion parameter σ^2 estimated in the model was 0.03793937. We have the same case as for $U - Be$, from the graphs we see that this model is not suitable.

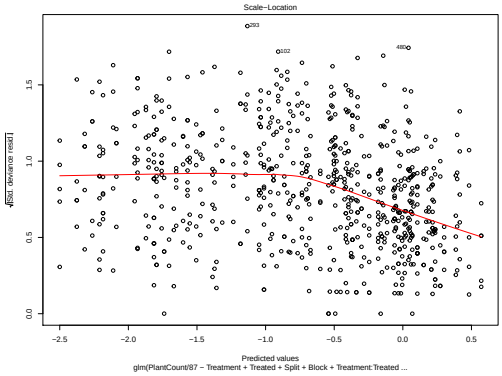
Table 6.9: Moorlands: Quasi with Log-Be

	Null deviance	Residual deviance
Selected model	72.338 on 719 df	22.858 on 576 df

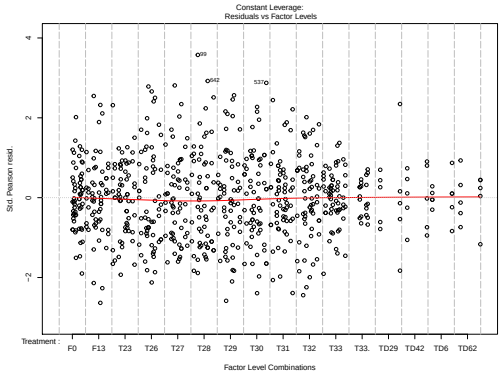


(a) Residuals vs. Fitted values

(b) Normal Q-Q plot



(c) Scale - Location graph



(d) Residuals vs Factor Levels graph

Figure 6.18: Moorlands - Graphical analysis of quasi GLM with Log - Be

Quasibinomial

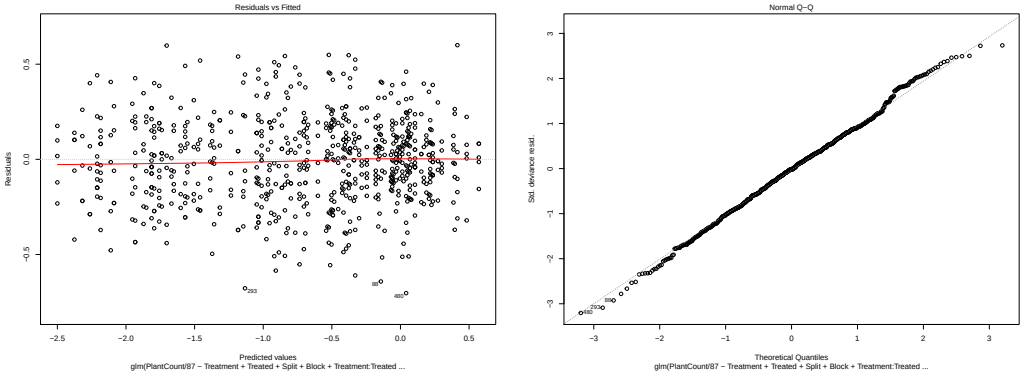
The very last distribution we tested on our model is quasibinomial, in which the variance of the response variables is assumed to be in the form $\sigma^2\mu(1-\mu)$. The response values need to be in the interval $\langle 0, 1\rangle$, therefore again as for binomial family we used the estimated parameter $n = 87$. Altering the initial model we obtain

$$\begin{aligned} PlantCount/87 \sim\ & Treatment + Treated + Split + Block \\ & + Treatment : Treated + Treatment : Split + Treatment : Block \\ & + Treated : Block + Split : Block + Treatment : Treated : Block \end{aligned}$$

with the dispersion parameter 0.06019964.

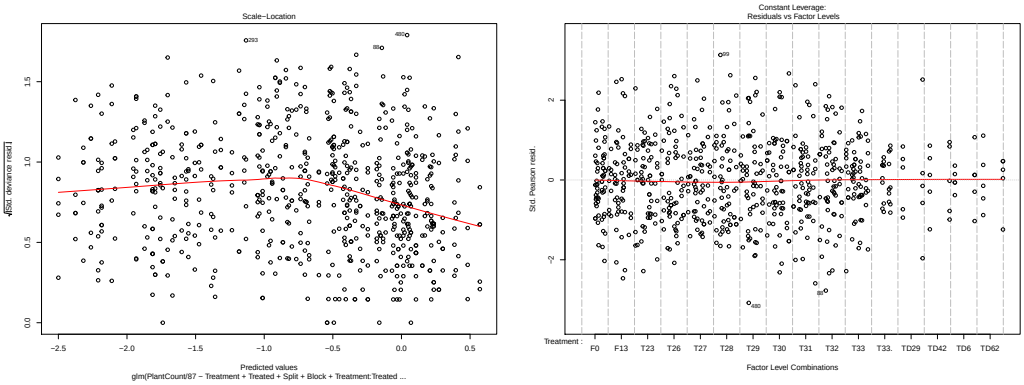
Table 6.10: Moorlands: Quasibinomial family

	Null deviance	Residual deviance
Selected model	114.580 on 719 df	35.885 on 576 df



(a) Residuals vs. Fitted values

(b) Normal Q-Q plot



(c) Scale - Location graph

(d) Residuals vs Factor Levels graph

Figure 6.19: Moorlands - Graphical analysis of Quasibinomial family

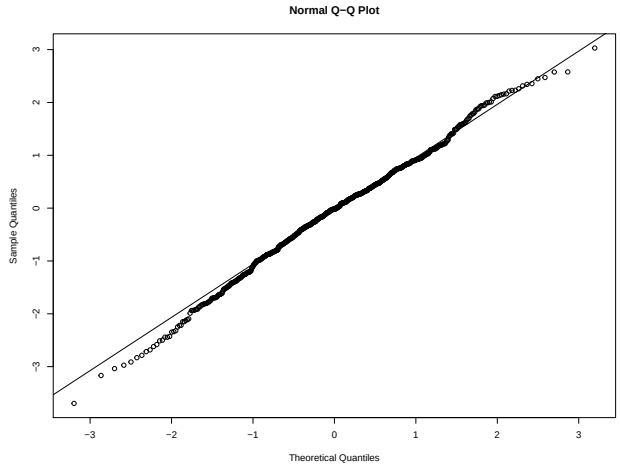


Figure 6.20: Moorlands - GLM with Quasibinomial family

6.4.2 Geranium

Discrete uniform - Bernoulli distribution

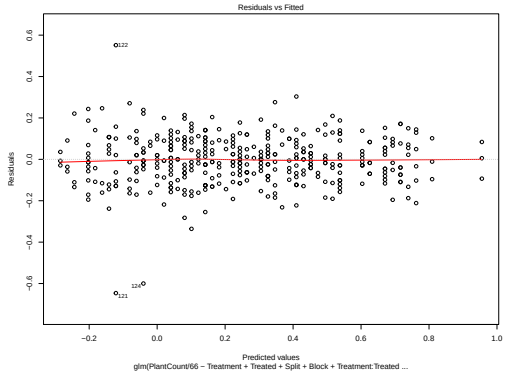
We assume the same properties as for Moorlands. We have chosen $a = 0$ and $b = 90$ and by altering the initial model we obtained

$$\begin{aligned} PlantCount/66 \sim\ & Treatment + Treated + Split + Block \\ & + Treatment : Treated + Treatment : Split + Treatment : Block \\ & + Treated : Block + Treatment : Treated : Block. \end{aligned}$$

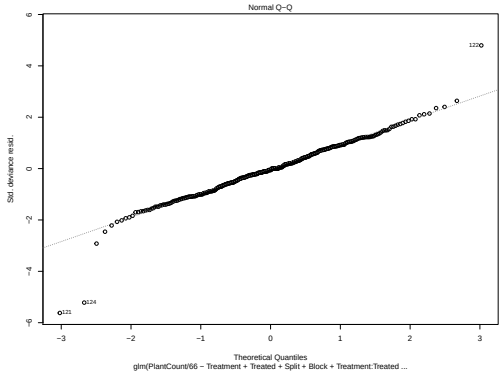
The dispersion parameter is 0.01983326.

Table 6.11: Geranium: Quasi with U-Be

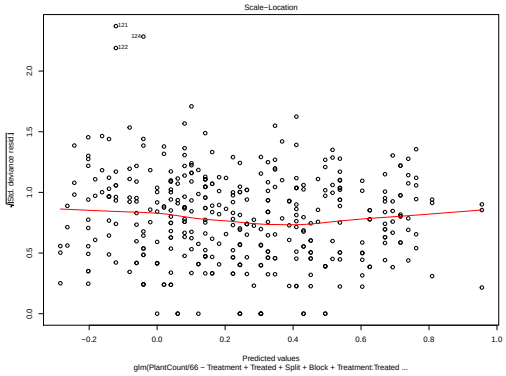
	Null deviance	Residual deviance
Selected model	8.1244 on 395 df	5.3998 on 264 df



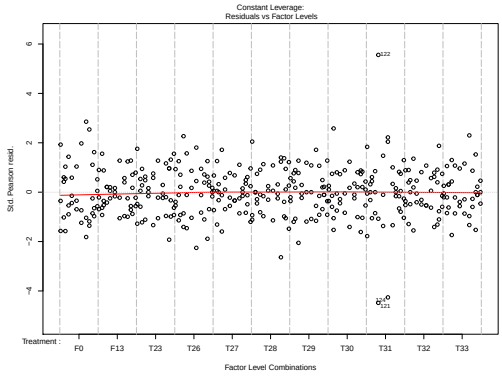
(a) Residuals vs. Fitted values



(b) Normal Q-Q plot



(c) Scale - Location graph



(d) Residuals vs Factor Levels graph

Figure 6.21: Geranium - Graphical analysis of quasi GLM with U - Be

Logarithmic - Bernoulli distribution

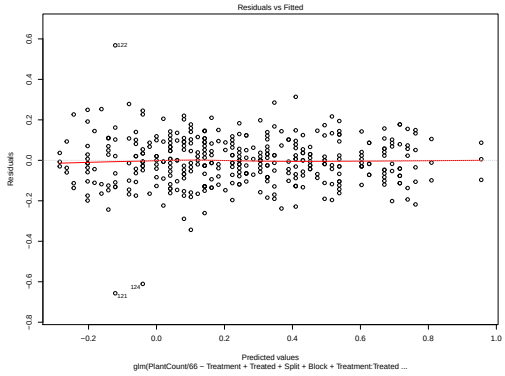
We assume the same properties as for Moorlands. We have chosen $q = 0.7$ and by altering the initial model we obtained

$$PlantCount/66 \sim Treatment + Treated + Split + Block + Treatment : Treated + Treatment : Split + Treatment : Block + Treated : Block + Treatment : Treated : Block.$$

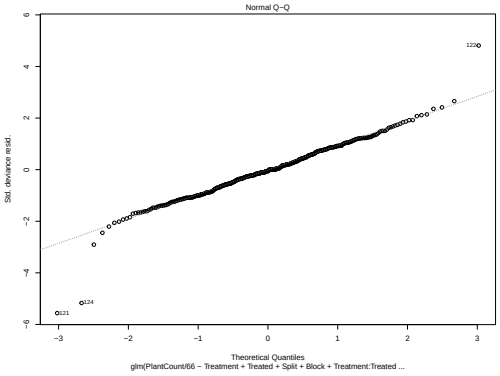
The dispersion parameter is 0.02093456.

Table 6.12: Geranium: Quasi with Log-Be

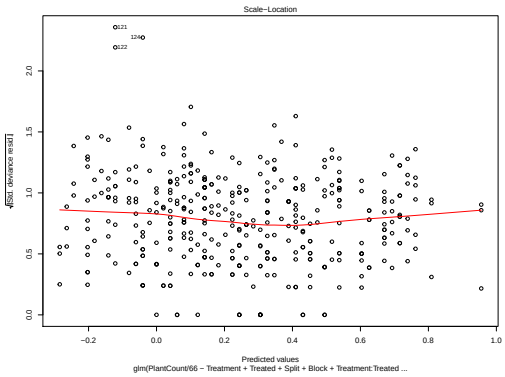
	Null deviance	Residual deviance
Selected model	8.5718 on 395 df	5.6902 on 264 df



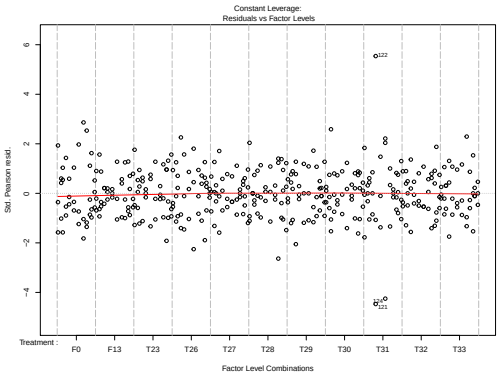
(a) Residuals vs. Fitted values



(b) Normal Q-Q plot



(c) Scale - Location graph



(d) Residuals vs Factor Levels graph

Figure 6.22: Geranium - Graphical analysis of quasi GLM with Log - Be

Quasibinomial

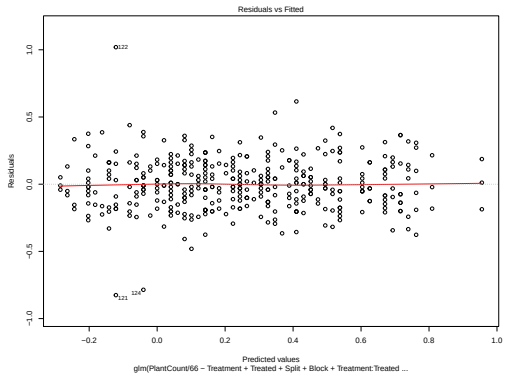
As for Moorlands, we used the estimated parameter $n = 66$. Altering the initial model we obtain

$$PlantCount/66 \sim Treatment + Treated + Split + Block + Treatment : Treated + Treatment : Split + Treatment : Block + Treated : Block + Treatment : Treated : Block$$

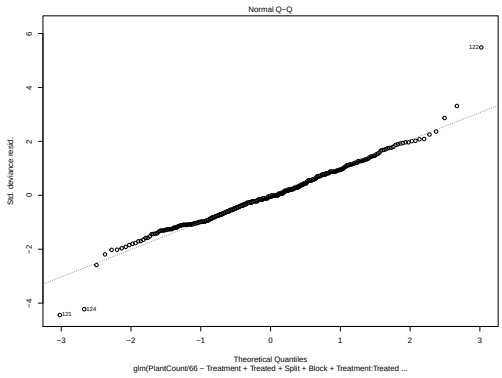
with the dispersion paramater 0.051838.

Table 6.13: Geranium: Quasibinomial family

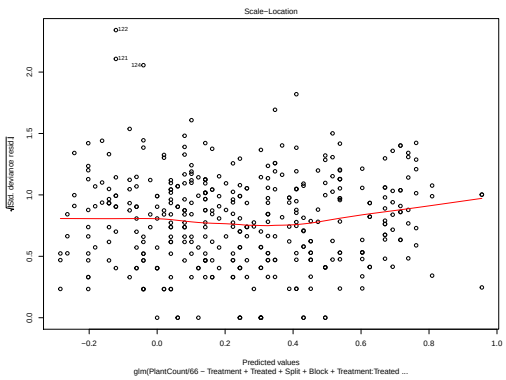
	Null deviance	Residual deviance
Selected model	21.649 on 395 df	14.267 on 264 df



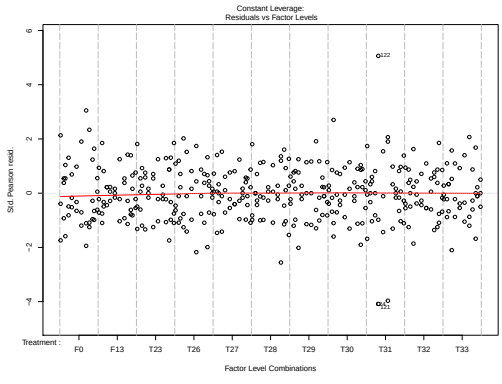
(a) Residuals vs. Fitted values



(b) Normal Q-Q plot



(c) Scale - Location graph



(d) Residuals vs Factor Levels graph

Figure 6.23: Geranium - Graphical analysis of Quasibinomial family

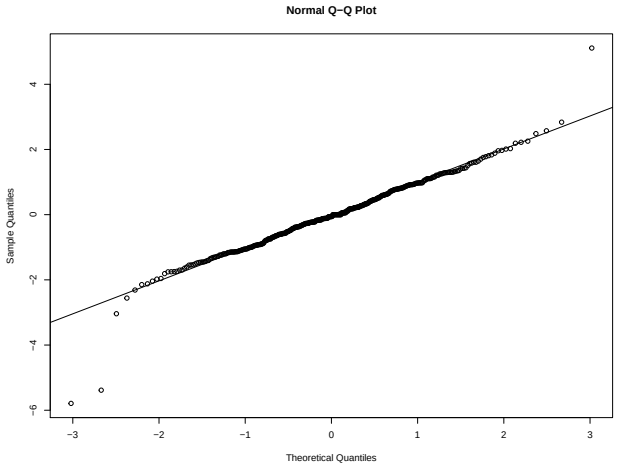


Figure 6.24: Geranium - GLM with Quasibinomial family

Conclusion

This master's thesis was dealing with analysis of discrete data, namely the number of plants that have grown in a field with the influence of the type of tillage and presence of fungicides. Around 50 seeds were sown by a seeding machine in five 1 meter long sections, therefore the number of seeds was random. This led us to introducing the term random sum S_N that is a sum of independent identically distributed random variables $X_i, i = 1, \ldots, n$, where their count N is a random variable as well.

Various properties of random sums were presented, such as expected value, variance, characteristic, probability generating, moment generating, and probability function. By multiple approaches for the case when variables $X_i \sim Be(p)$, we have shown that if the variable N has binomial, Poisson, or negative binomial distribution, then the distribution of random sum S_N is preserved.

We described the maximum likelihood method for estimating parameters of distribution with its principles, we defined the term exponential family of distributions and derived the score and Fisher information. For binomial, Poisson, and negative binomial distribution finding the estimator of parameters is simple process as they belong to the exponential family. On the other hand, the mixture of discrete uniform and Bernoulli distribution has to be solved numerically and mixture of logarithmic and Bernoulli is called the logarithmic-with-zeros distribution and its parameters can be transformed so we can use the ML method.

Models with exponential family of distributions can be fitted with generalized linear models, which analyze parallel situations to linear regression models, where the response variable is not necessarily Normal. We briefly discuss the quasilikelihood method, which can be used for the rest of the models if we know the relationship between expected value and variance.

Lastly we analyzed data from two separate fields in Australia in R. From the overall graphical analysis we couldn't make any strong conclusions, but we see that for Geranium the mean is around 40, whereas for Moorlands due to a lot of values under 20 the mean might be lower. The difference in the

side of the seeder for Moorlands could be assessed from the $Treatment = F0$. We computed the mean and variance by variable *Plot* and compared with the relationships we derived in the second chapter, however, the graphs didn't resemble the theoretical ones.

For model fitting, we took into consideration the binomial, Poisson, negative binomial, quasi with $U - Be$ and $Log - Be$, and quasibinomial family. The most suitable model for Moorlands field was the one with binomial family, as can be seen from the graphical analysis, as well as from the value of residual deviance 35.885 on 576 df. For Geranium the most suitable model was also with binomial family, its residual deviance was 14.267 on 264 df.

Bibliography

[1] AGRESTI, Alan. *Categorical Data Analysis*. Gainesville: John Wiley, 2002. ISBN 0-471-36093-7.

[2] ANDĚL, Jiří. *Základy matematické statistiky*. 4th ed. Praha: Matfyzpress, 2013. ISBN 978-80-7378-162-0.

[3] DOBSON, Annette J. and BARNETT, A. *An Introduction to Generalized Linear Models*. 3rd ed. Boca Raton: Chapman & Hall/CRC, 2002. ISBN 1-58488-165-8.

[4] FELLER, William. *An Introduction to Probability Theory and Its Applications*. New York: John Wiley, 1950.

[5] GÜNEL, Erdogan and CHILKO, Daniel. Estimation of Parameter n of the Binomial Distribution. *Communications in Statistics - Simulation and Computation*. 1989, vol. 18, p. 537-551.

[6] HRDLIČKOVÁ, Zuzana. *Log-lineární modely s Poissonovskými proměnnými*. Brno: Masarykova univerzita, přírodovědecká fakulta, 2002.

[7] JOHNSON, Norman L., KOTZ, Samuel, and KEMP, Adrienne W. *Univariate Discrete Distributions*. New York: Wiley, 1992.

[8] DUPAČ, Václav and DUPAČOVÁ, Jitka. *Markovovy procesy*. 1. [Part]. Praha: Statní pedagogické nakladatelství, 1975.

[9] KHATRI, C. G. On the Distributions Obtained by Varying the Number of Trials in a Binomial Distribution. *Annals of the Institute of Statistical Mathematics*. 1961, vol. 13, p. 47-51. ISSN 1572-9052.

[10] MCCULLAGH, P. and NELDER, J. A. *Generalized Linear Models*. 2nd ed. Chapman & Hall/CRC, 1989. ISBN 978-0-41-231760-6.

[11] RÉNYI, Alfréd. *Teorie pravděpodobosti: vysokoškolská učebnice ČSR*. Praha: Academia, 1972.

[12] ROSS, Sheldon M. *Introduction to Probability Models*. Amsterdam: Academic Press, 2010. ISBN 978-0-12-375686-2.

[13] WIMMER, Gejza. *Diskrétne jednorozmerné rozdelenia pravdepodobnosti*. Praha: Matfyzpress, 2000. ISBN 80-85863-60-X.

[14] WITKOWSKY, Viktor, 2018. *The Characteristic Function Toolbox*. [online]. [Cit. 20.4.2018]. Available from https://github.com/witkovsky/CharFunTool.

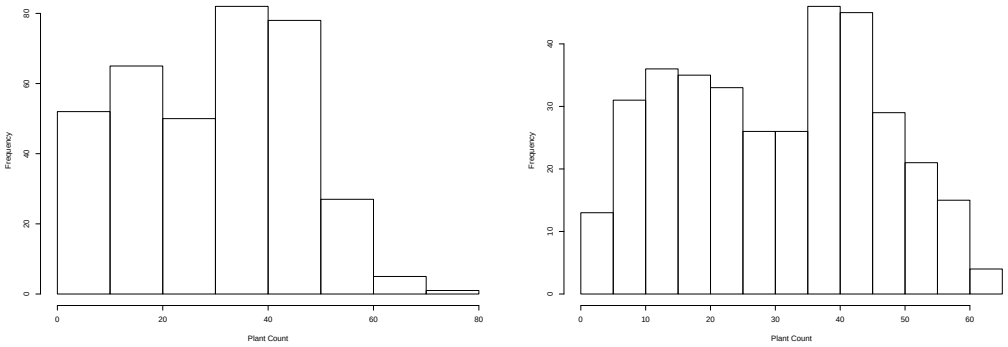
Lists of symbols and abbreviations

Object	Name
$f^{(a)}$	a-th derivative of the function f
$\mathbb{N}$	set of natural numbers
$\mathrm{E}[X]$	Expected value of the random variable X
$\mathrm{var}[X]$	Variance of the random variable X
$P(X = x)$	Probability of a random variable X equal to x
$\psi_X(t)$	Characteristic function of a random variable X
$G_X(s)$	Probability generating function of a random variable X
$M_X(t)$	Moment generating function of a random variable X
cf	Characteristic function
pdf	Probability density function
pgf	Probability generating function
mgf	Moment generating function
$Bi(n, p)$	Binomial distribution with parameters n and p
$Be(p)$	Bernoulli distribution with parameter p
$Po(\lambda)$	Poisson distribution with parameter λ
$NBi(\kappa, p)$	Negative binomial distribution with parameters κ and p
$U(a, b)$	Discrete uniform distribution with parameters a and b
$Log(p)$	Logarithmic distribution with parameter p

Object	Name
GLM	Generalized linear model
ML	Maximum likelihood
df	Degrees of Freedom
χ^2	Chi-squared distribution

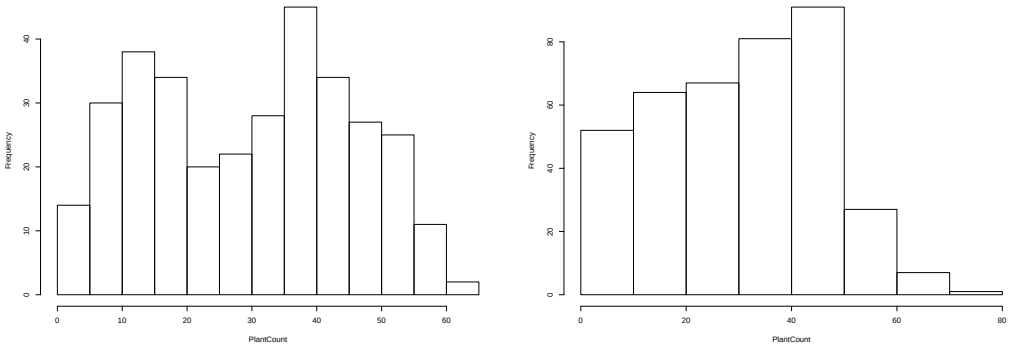
Appendix

Moorlands



(a) Split = N (b) Split = S

Figure A.1: Moorlands - Histograms of PlantCount



(a) Treated = Treated (b) Treated = Untreated

Figure A.2: Moorlands - Histograms of PlantCount by Treated

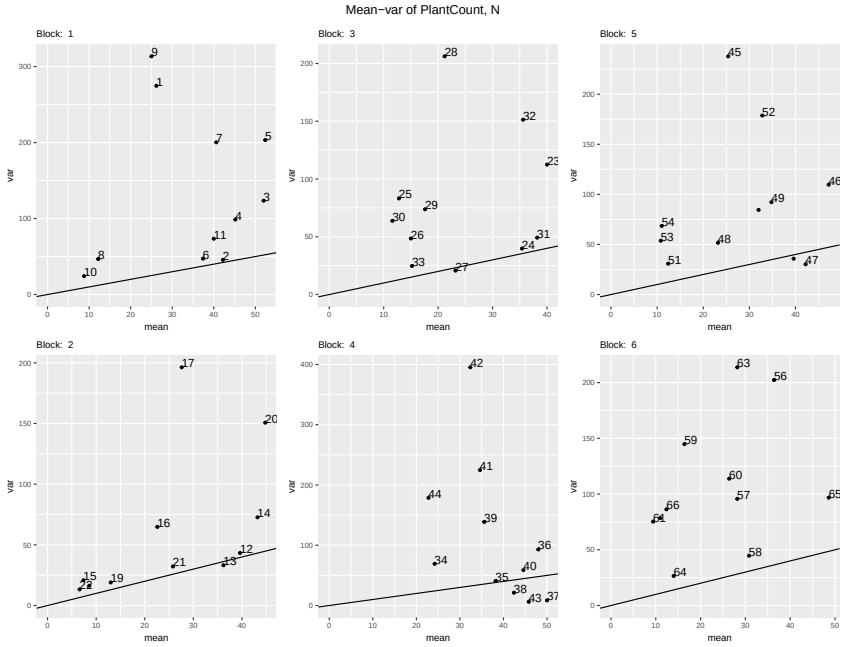


Figure A.3: Moorlands - Block with Split = N

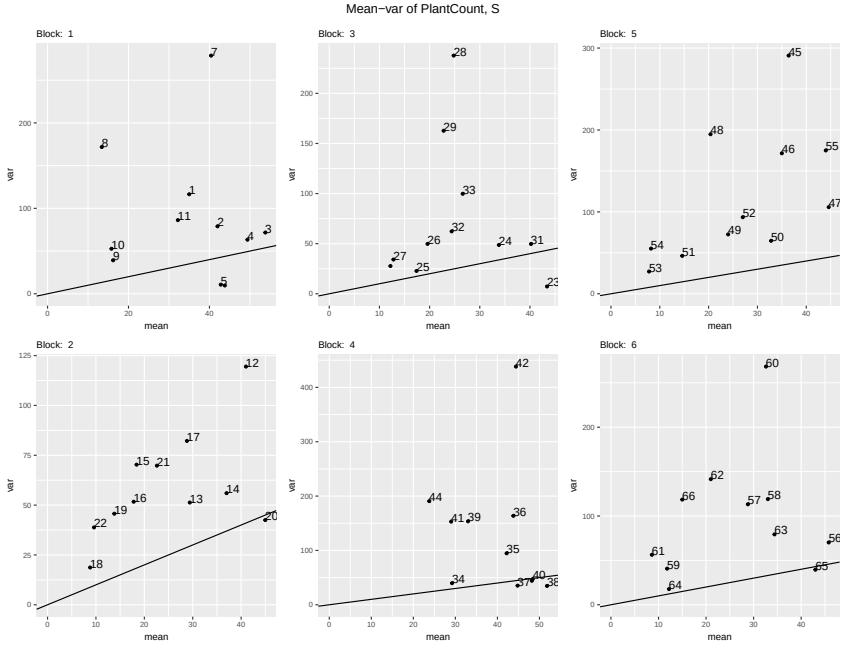


Figure A.4: Moorlands - Block with Split = S

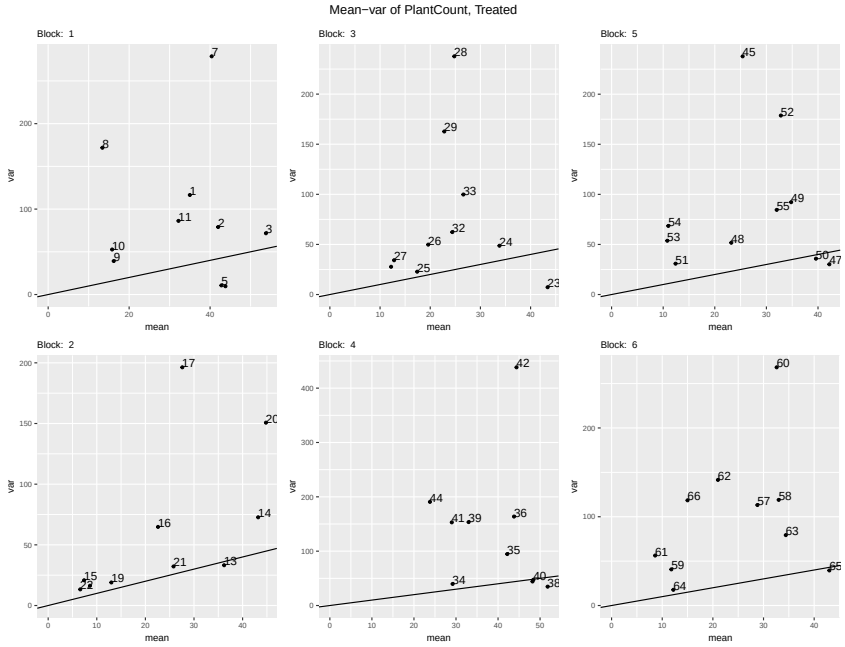


Figure A.5: Moorlands - Block with Treated = Treated

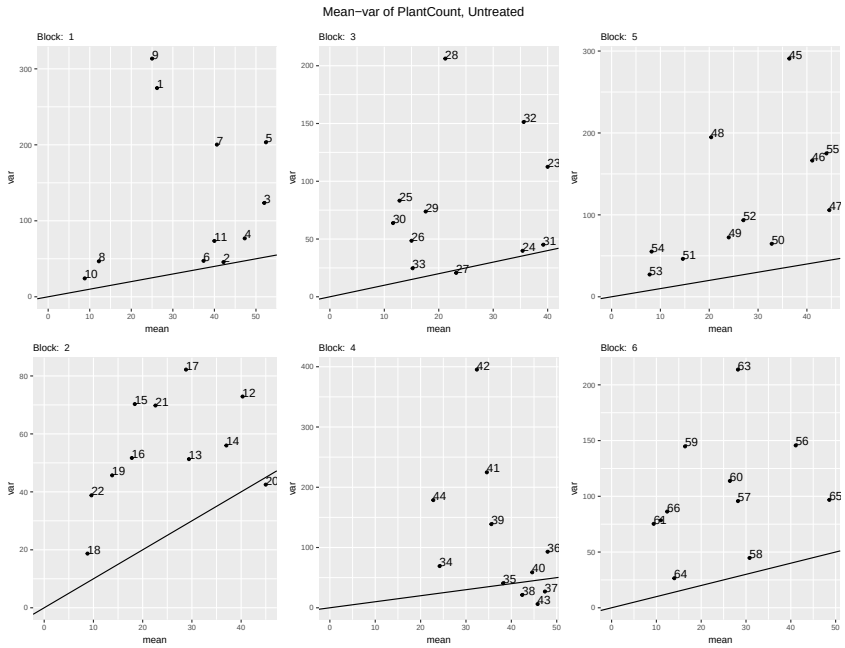


Figure A.6: Moorlands - Block with Treated = Untreated

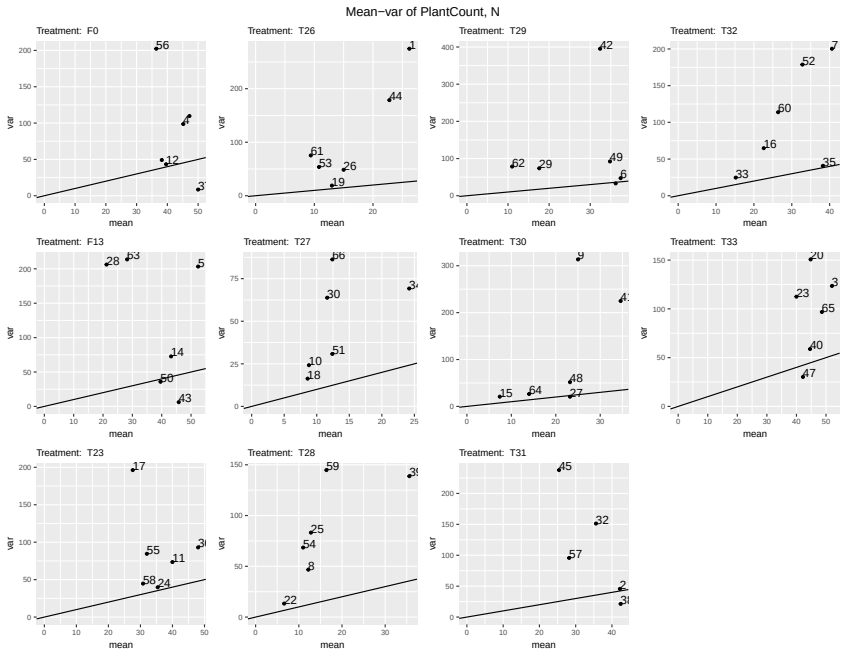


Figure A.7: Moorlands - Treatment with Split = N

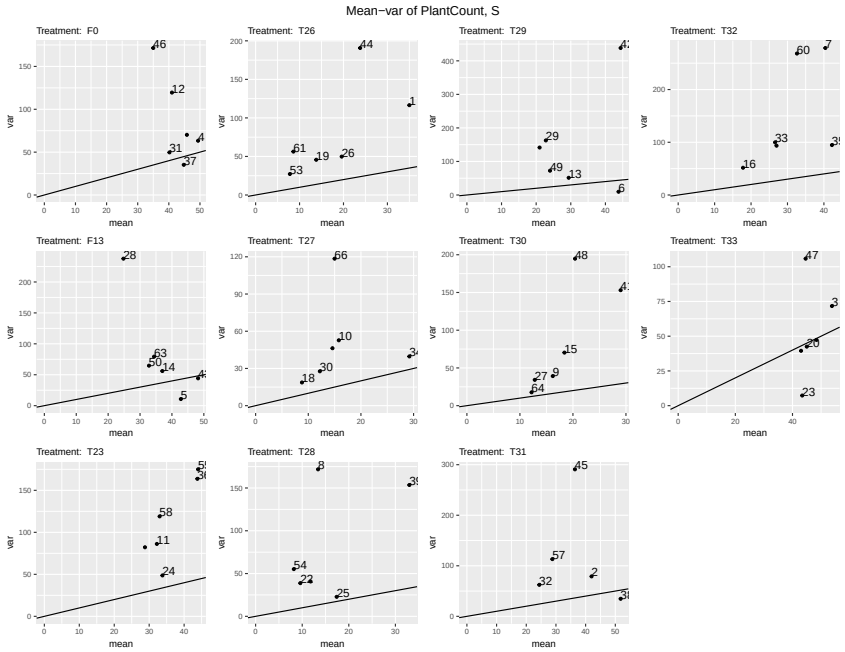


Figure A.8: Moorlands - Treatment with Split = S

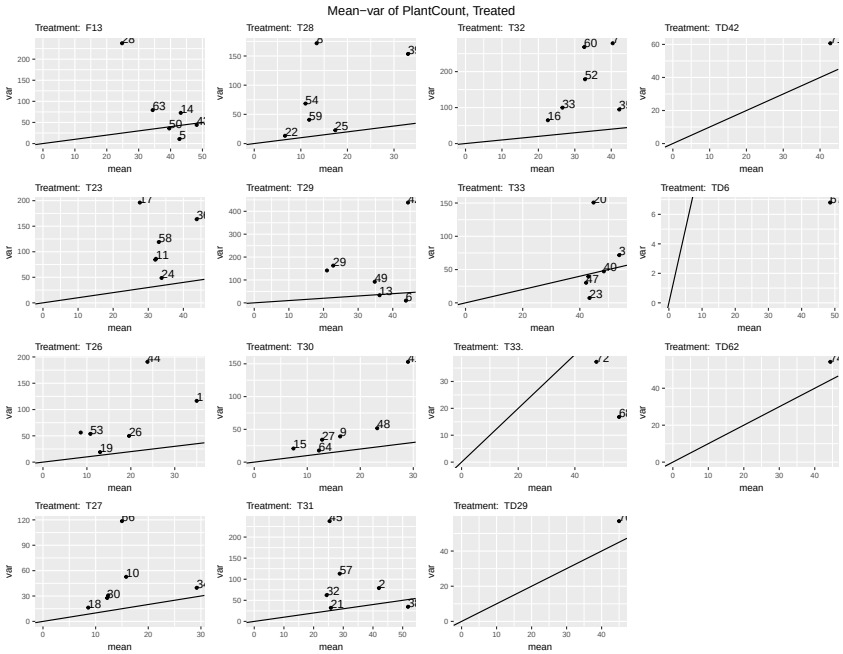


Figure A.9: Moorlands - Treatment with Treated = Treated

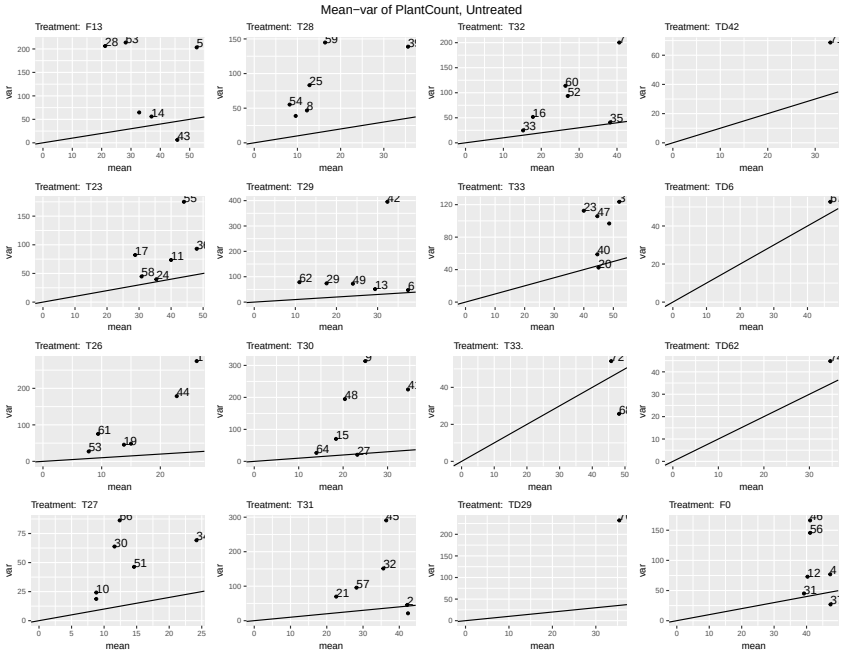
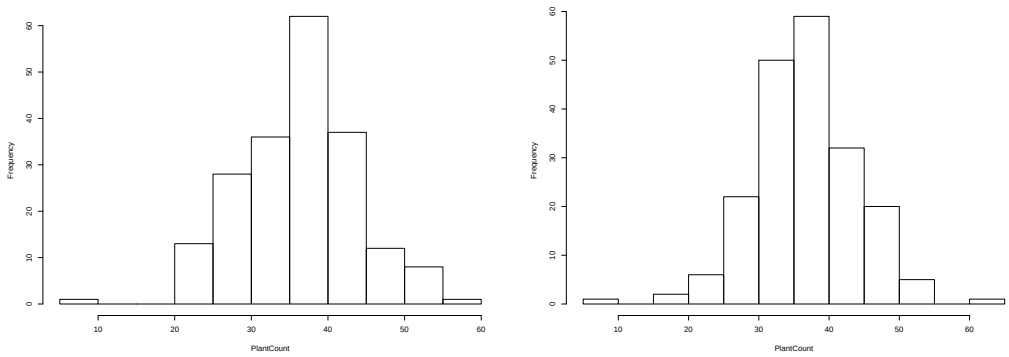


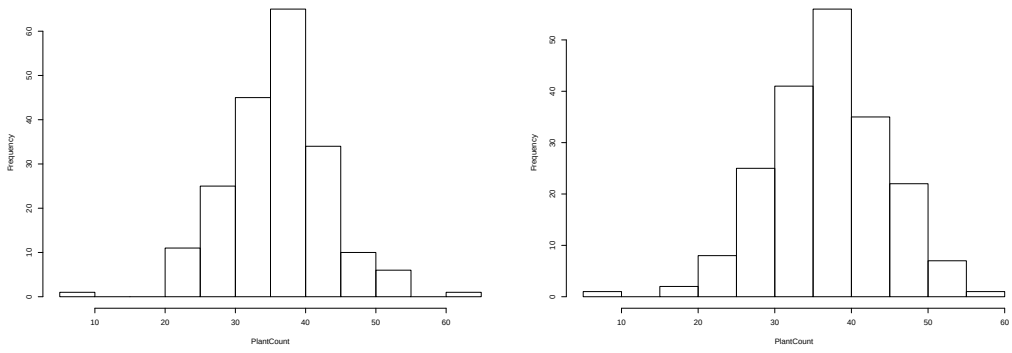
Figure A.10: Moorlands - Treatment with Treated = Untreated

Geranium



(a) Split = N (b) Split = S

Figure A.11: Geranium - Histograms of PlantCount by Split



(a) Treated = Treated (b) Treated = Untreated

Figure A.12: Geranium - Histograms of PlantCount by Treated

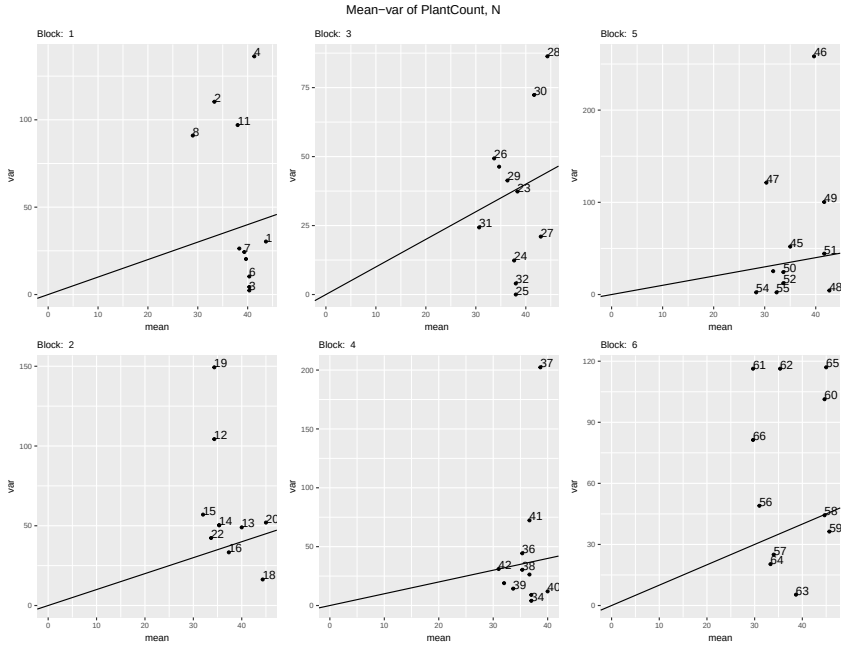


Figure A.13: Geranium - Block with Split = N

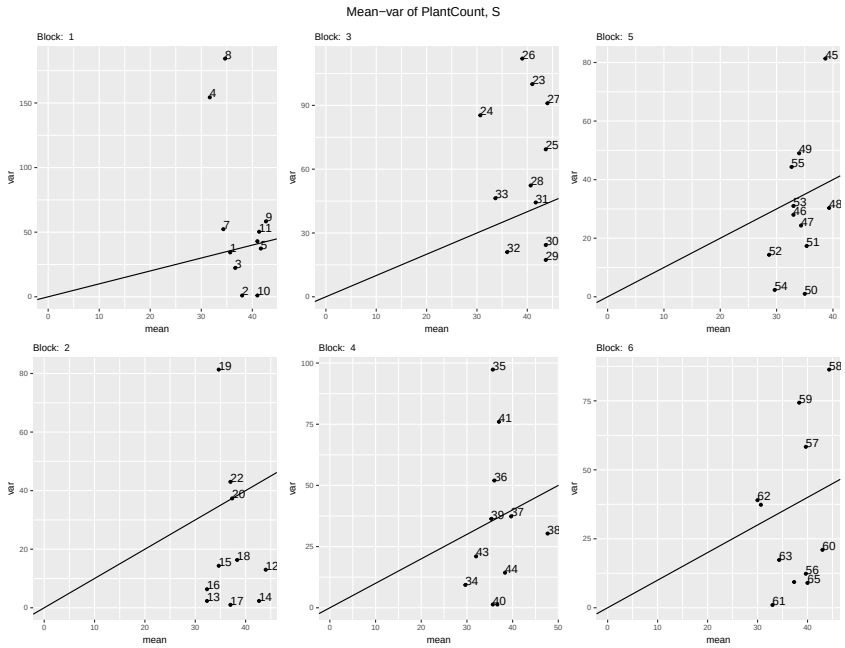


Figure A.14: Geranium - Block with Split = S

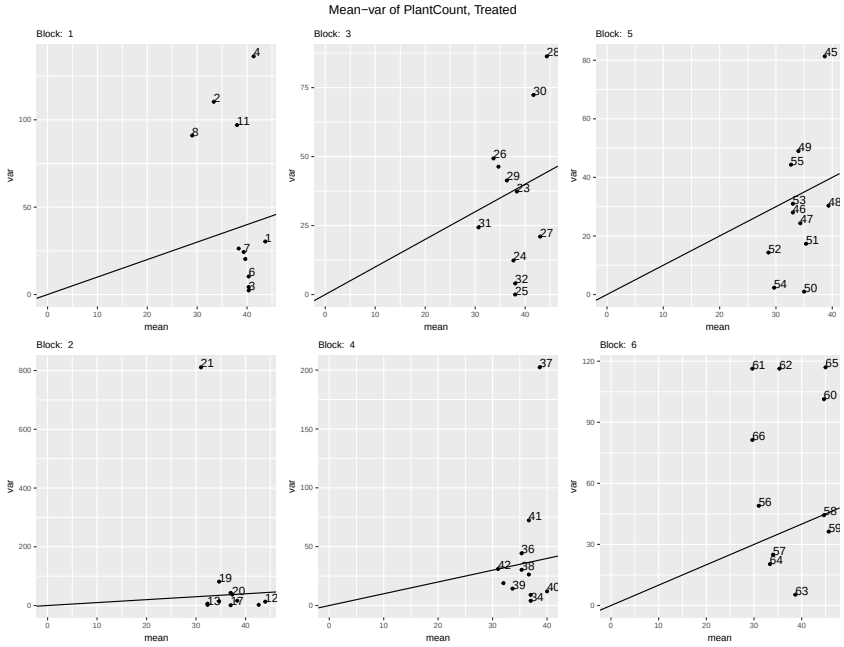


Figure A.15: Geranium - Block with Treated = Treated

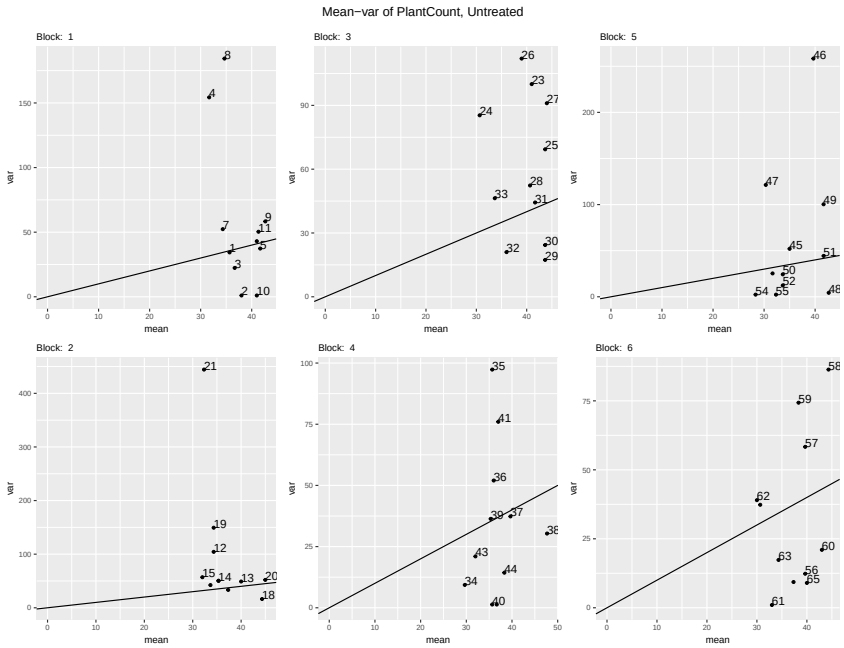


Figure A.16: Geranium - Block with Treated = Untreated

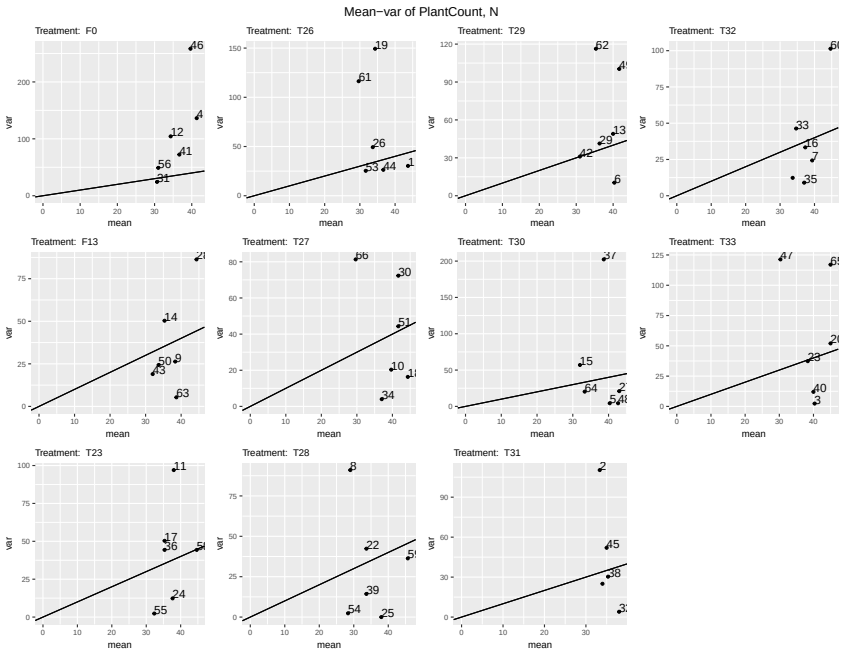


Figure A.17: Geranium - Treatment with Split = N

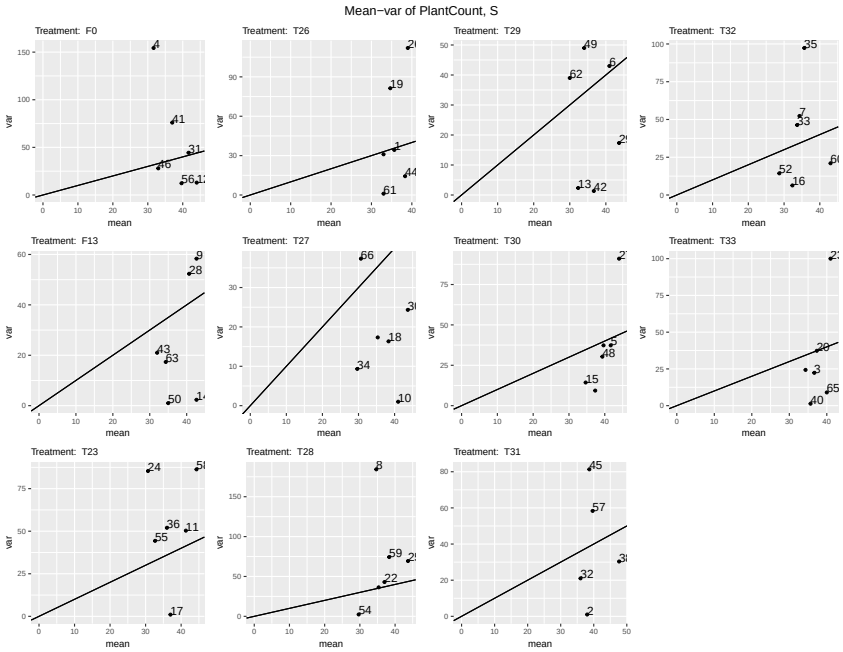


Figure A.18: Geranium - Treatment with Split = S

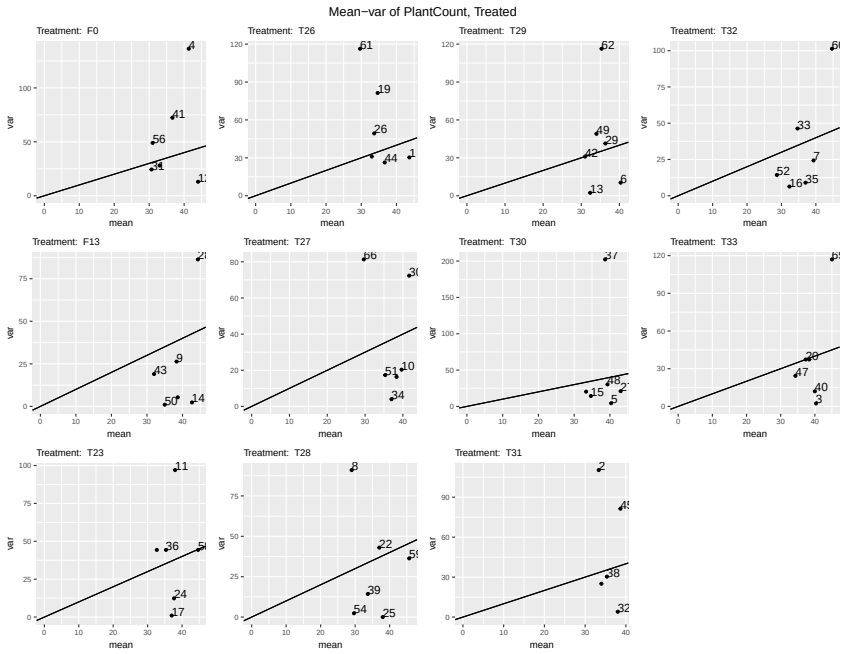


Figure A.19: Geranium - Treatment with Treated = Treated

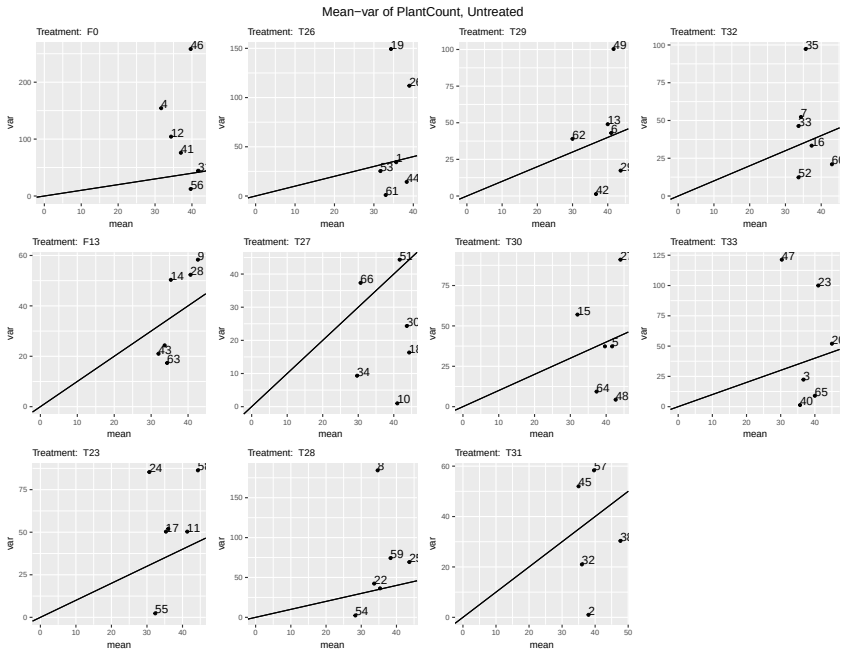


Figure A.20: Geranium - Treatment with Treated = Untreated

Electronic Appendix Index

1. `moorlands.r` - Graphical analysis for Moorlands
2. `geranium.r` - Graphical analysis for Geranium
3. `quasi2.r` - function `quasi2`
4. `moorlands-glm.r` - GLM analysis for Moorlands
5. `geranium-glm.r` - GLM analysis for Geranium